Struct Multidisc Optim (2017) 55:1029–1050
DOI 10.1007/s00158-016-1550-y

RESEARCH PAPER

Remarks on multi-fidelity surrogates

Chanyoung Park[1] · **Raphael T. Haftka**[1] · **Nam H. Kim**[1]

Received: 12 November 2015 / Revised: 28 May 2016 / Accepted: 25 July 2016 / Published online: 16 August 2016

Abstract Different multi-fidelity surrogate (MFS) frameworks have been used for optimization or uncertainty quantification. This paper investigates differences between various MFS frameworks with the aid of examples including algebraic functions and a borehole example. These MFS include three Bayesian frameworks using 1) a model discrepancy function, 2) low fidelity model calibration and 3) a comprehensive approach combining both. Three counterparts in simple frameworks are also included, which have the same functional form but can be built with ready-made surrogates. The sensitivity of frameworks to the choice of design of experiments (DOE) is investigated by repeating calculations with 100 different DOEs. Computational cost savings and accuracy improvement over a single fidelity surrogate model are investigated as a function of the ratio of the sampling costs between low and high fidelity simulations. For the examples considered, MFS frameworks were found to be more useful for saving computational time rather than improving accuracy. For the Hartmann 6 function example, the maximum cost saving for the same accuracy was 86 %, while the maximum accuracy improvement for the same cost was 51 %. It was also found that DOE can substantially change the relative standing of different frameworks. The cross-validation error appears to be a reasonable candidate for estimating poor MFS frameworks for a specific problem but it does not perform well compared to choosing single fidelity surrogates.

Keywords Multi-fidelity surrogate · Bayesian · Calibration · Discrepancy function

✉ Chanyoung Park
cy.park@ufl.edu

[1] Department of Mechanical and Aerospace Engineering, University of Florida, PO Box 116250, Gainesville, FL 32611-6250, USA

Nomenclature

$\boldsymbol{\delta}$	A discrepancy data set for given ρ ($\boldsymbol{\delta} = \mathbf{y}_H - \rho\mathbf{y}_L^c$).
$\delta(\mathbf{x})$	An unknown true value of a discrepancy function value at $\mathbf{x}$.
$\hat{\delta}(\mathbf{x})$	A predictor for a discrepancy function value at $\mathbf{x}$.
$\Delta(\mathbf{x})$	A prior model (GP model) for predicting a discrepancy function value at $\mathbf{x}$ for Bayesian MFS frameworks. Note that a discrepancy function can be a function for given ρ.
$\Delta(\mathbf{x})\|\boldsymbol{\delta}$	An updated discrepancy function model with a discrepancy data set.
$\boldsymbol{\lambda}$	A roughness parameter vector.
$\boldsymbol{\theta}$	A calibrated parameter vector (a constant vector).
ρ	A scalar for a low fidelity function.
σ	A process standard deviation.
$\boldsymbol{\xi}(\mathbf{x})$	A vector of shape functions.
$\mathbf{b}$	A coefficient vector (a constant vector).
$\mathbf{q}$	A calibration variable vector (a variable vector).
$\mathbf{x}$	An input variable vector (a variable vector).
$\mathbf{y}$	A data set.
$y(\mathbf{x})$	An unknown true function value at $\mathbf{x}$.
$\hat{y}(\mathbf{x})$	A surrogate predictor for a function value at $\mathbf{x}$.
$Y(\mathbf{x})$	A prior model for predicting a function at $\mathbf{x}$ for Bayesian frameworks and Kriging surrogate. A prior model is a fitted GP model ($Z(\mathbf{x})$) parameterized with a linear polynomial trend function, which approximates the true function, and the corresponding uncertainty in the trend function. The parameters of a prior model are found by samples for maximum consistency.

$Y(\mathbf{x})\|\mathbf{y}$	An updated model with a data set. The trend function and the corresponding uncertainty of a prior model are updated with samples.
$\mathbf{y}_H$	A high fidelity data set.
$\mathbf{y}_h^i$	The i-th data point of a high fidelity data set.
$y_H(\mathbf{x})$	An unknown true high fidelity function value at $\mathbf{x}$.
$\hat{y}_H(\mathbf{x})$	An MFS predictor for a high fidelity function value at $\mathbf{x}$.
$Y_H(\mathbf{x})$	A prior model (GP model) for predicting a high fidelity function value at $\mathbf{x}$ for Bayesian MFS frameworks. This model can be a linear combination of a low fidelity model and a discrepancy function model.
$Y_H(\mathbf{x})\|\mathbf{y}_H, \mathbf{y}_L$	An updated high fidelity model with low and high fidelity data sets.
$\mathbf{y}_L$	A low fidelity data set.
$\mathbf{y}_L^c$	A low fidelity data set at locations common to those of high fidelity data points.
$\mathbf{y}_l^i$	The i-th data point of a low fidelity data set.
$y_L(\mathbf{x})$	An unknown true low fidelity function value at $\mathbf{x}$.
$\hat{y}_L(\mathbf{x})$	An MFS predictor for a low fidelity function value at $\mathbf{x}$.
$\hat{y}_L(\mathbf{x}, \boldsymbol{\theta})$	An MFS predictor for a low fidelity function value at $\mathbf{x}$ for a given calibrated parameter vector $\boldsymbol{\theta}$.
$Y_L(\mathbf{x})$	A prior model (GP model) for predicting a low fidelity function value at $\mathbf{x}$ for Bayesian MFS frameworks.
$Y_L(\mathbf{x}, \boldsymbol{\theta})$	A prior model (GP model) for a prediction of a low fidelity function value at $\mathbf{x}$ for a given calibrated parameter vector $\boldsymbol{\theta}$.
$Y_L(\mathbf{x})\|\mathbf{y}_L$	An updated low fidelity model with a low fidelity data set.

1 Introduction

Surrogate models, also known as meta-models, have been used as a cheap approximate model, which can be built with several dozens of samples. However, for many high fidelity simulations, the cost for obtaining enough number of samples for achieving reasonable accuracy is high. Multi-fidelity surrogate (MFS) models have been developed to compensate for expensive high fidelity samples with cheap low fidelity samples. Although, several Gaussian process (GP) based Bayesian MFS frameworks have been introduced, the benefits of the Bayesian frameworks over simple frameworks have been rarely studied. In addition, the performances of different frameworks have been rarely compared, which is the main focus of this paper.

MFS frameworks based on a discrepancy function are built by combining a low-fidelity simulation (or surrogate) with a discrepancy surrogate, which models the difference between low and high fidelity sample sets. Discrepancy-based MFS frameworks have been used in design optimization to alleviate computational burden. For example, Balabanov et al. (1998) used linear regression to combine coarse and fine finite element models, while Mason et al. (1998) used 2D and 3D finite element models as a low and high fidelity model, respectively, for aircraft structural optimization. The same approach was used to combine aerodynamic prediction from a cheap linear theory with expensive Euler solutions for aircraft aerodynamic optimization (Knill et al. 1999). A Bayesian discrepancy-based MFS using GP was introduced by Kennedy and O'Hagan (2000). The Bayesian model allows to incorporate prior information (Kennedy and O'Hagan 2000; Qian and Wu 2008). Co-Kriging (Sacks et al. 1989; Lophaven et al. 2002) provides an equivalent surrogate to the Bayesian formulation with a non-informative prior and has good computational characteristics (Forrester et al. 2007; Kuya et al. 2011; Han et al. 2012; Le Gratiet 2013). Note that the discrepancy function based MFS frameworks can handle data from more than two fidelities.

Model calibration is another strategy for building MFS by fitting a low fidelity surrogate with tuned parameters which improve agreement between the surrogate and high fidelity sample set. A simple framework is to find parameters that minimize the discrepancy between a calibrated low fidelity surrogate and high fidelity sample set (Zheng et al. 2013). GP-based Bayesian calibration frameworks were also introduced (Kennedy and O'Hagan 2001a; Higdon et al. 2004; Bayarri et al. 2007; McFarland et al. 2008). The Bayesian frameworks find the best calibration parameters that are the most statistically consistent with high fidelity samples (McFarland et al. 2008; Prudencio and Schulz 2012). A comprehensive Bayesian MFS model that uses both calibration and discrepancy was proposed by Kennedy and O'Hagan (2001a) offering greater flexibility, although this is the most complex framework.

The objectives of this paper are: (1) to review the characteristics and differences of GP-based Bayesian and simple MFS frameworks, (2) to investigate the performance of MFS frameworks in terms of accuracy and predictability of error, and (3) to investigate the performance of prediction sum of squares (*PRESS*) based on cross validation errors as a surrogate performance estimator. The paper is organized as follows. Section 2 presents MFS frameworks used in this paper. Section 3 describes the methodology of the investigation and metrics. Section 4 presents numerical examples, followed by discussions in Section 5. Each section is intended to be self-explanatory such that skipping earlier sections does not hamper reading later sections.

2 Multi-fidelity surrogate frameworks

The first objective of this paper is to review the differences between MFS frameworks based on three commonly used approaches: using (1) a model discrepancy function, (2) a low fidelity model parameter calibration and (3) a combination approach. We considered three simple MFS frameworks and three Bayesian frameworks using those approaches. The simple frameworks refer to frameworks that build an MFS using any surrogate. For example, Balabanov et al. (1998) used a polynomial response surface MFS that is the sum of a low fidelity surrogate and a model discrepancy surrogate. The characteristics, assumptions of the considered frameworks and their differences will be described in the following subsections.

All Bayesian MFS frameworks are based on Gaussian Process (GP) for modeling their prediction uncertainties. Alternatively, we have simple frameworks that construct an MFS with regular single fidelity surrogates. In order to minimize the effect of the choice of surrogate for the simple frameworks, we use Kriging surrogates, which are also based on GP. Furthermore, the combination of Kriging and the simple discrepancy framework is a special case of the Bayesian discrepancy framework.

The Bayesian variant of Kriging shares many features for modeling prediction uncertainty and fitting processes with the Bayesian MFS. We therefore start by describing the Bayesian Kriging surrogate so as to help understanding the Bayesian MFS frameworks.

2.1 Bayesian Kriging surrogate

Kriging surrogate is a well-known generalized linear regression model based on GP (Sacks et al. 1989; Martin and Simpson 2005; Lophaven et al. 2002; and Ryu et al. 2002). Not surprisingly, there is a parallel effort to derive a Kriging surrogate model using Bayesian techniques. The mathematical form of Bayesian Kriging surrogate is identical to that of the generalized linear regression model for non-informative prior (Lophaven et al. 2002; O'Hagan 1992). Bayesian Kriging surrogate provides a prediction with errors in a form of t distribution (O'Hagan 1992). The mean and standard deviation of the t distribution represent the Kriging predictor and the prediction uncertainty, respectively. The Kriging predictor is expressed as

$$\hat{y}(\mathbf{x}) = E\left(Y(\mathbf{x})\middle|\mathbf{y}\right) \tag{1}$$

where $Y(\mathbf{x})$ is a prior GP model and $Y(\mathbf{x})|\mathbf{y}$ is the updated model with sample set $\mathbf{y}$.

Figure 1a illustrates a prior GP model with a quadratic trend function $\hat{y}(\mathbf{x})$, which is the mean function, and two standard deviation intervals (blue-colored region). The prior model characterizes the trend and prediction uncertainty, where hyper parameters are estimated for the best consistency of the GP model with samples. Figure 1b shows a Kriging prediction and two standard deviation intervals obtained by updating the GP model $Y(\mathbf{x})$ in Fig. 1a with sample set $\mathbf{y}$.

A GP model typically includes a linear combination of shape functions term $\boldsymbol{\xi}(\mathbf{x})^{\mathrm{T}}\mathbf{b}$ as a trend function and a normal random variable $Z(\mathbf{x}) \sim N(0, \sigma^2)$, which models the prediction uncertainty of the trend function, as

$$Y(\mathbf{x}) = \boldsymbol{\xi}(\mathbf{x})^{\mathrm{T}}\mathbf{b} + Z(\mathbf{x}) \tag{2}$$

where $\boldsymbol{\xi}(\mathbf{x})$ is a given vector of shape functions (e.g. polynomials), and $\mathbf{b}$ is a coefficient vector. The prediction uncertainties at two points $\mathbf{x}$ and $\mathbf{x}'$ are assumed to be correlated via a covariance function. A common covariance function is

$$\mathrm{cov}\left(Z(\mathbf{x}), Z\left(\mathbf{x}'\right)\right) = \sigma^2 \exp\left(-\left(\mathbf{x}-\mathbf{x}'\right)^{\mathrm{T}}\boldsymbol{\Omega}\left(\mathbf{x}-\mathbf{x}'\right)\right) \tag{3}$$

where $\boldsymbol{\Omega}$ is a diagonal matrix $\boldsymbol{\Omega} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ denotes a diagonal matrix with n elements. For example, the (1,1) element of $\boldsymbol{\Omega}$ is $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}^{\mathrm{T}}$ is a vector of hyper parameters defining roughness via regulating the range of correlation for n dimensional $\mathbf{x}$, and σ is the process standard deviation determining the magnitude of the error.

Fitting a Kriging surrogate $\hat{y}(\mathbf{x})$ has two steps: (1) a prior GP model is fitted to samples by estimating hyper parameters $\{\mathbf{b}, \sigma, \boldsymbol{\lambda}\}$ using the maximum a posteriori estimation which takes a mode of the posterior distribution for the best hyper parameter estimates, and (2) the prior GP model is updated with samples using the formula of conditional distributions for a multivariate normal distribution. For details about the processes, readers are referred to O'Hagan (1992) and Rasmussen (2004).

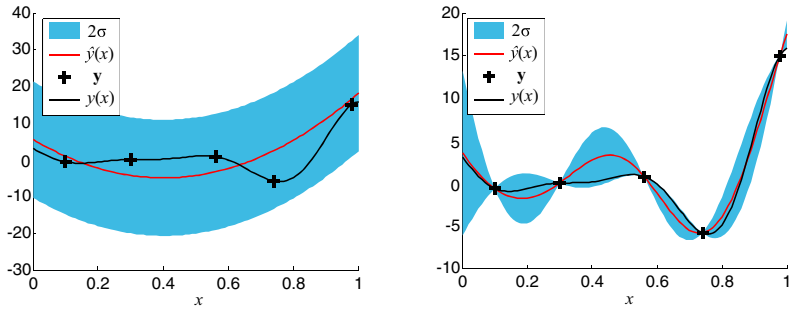
2.2 Discrepancy function based frameworks

2.2.1 Simple discrepancy MFS framework

This framework provides a convenient way of fitting an MFS with any surrogate model. The MFS employs two surrogates, $\hat{y}_L(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$, which approximate the low fidelity function and the discrepancy, respectively, as

$$\hat{y}_H(\mathbf{x}) = \rho\hat{y}_L(\mathbf{x}) + \hat{\delta}(\mathbf{x}) \tag{4}$$

where ρ is a regression scalar minimizing the discrepancy between the scaled low fidelity surrogate $\rho\hat{y}_L(\mathbf{x})$ and the high fidelity sample set at the common sampling points, where sample locations for high and low fidelities are the same. Note that modeling ρ as a function for inputs is an active research area (Fischer and Grandhi 2014, 2015).



(a) Estimating the trend function and hyper parameters for a prior GP model using samples

(b) Updating the prior GP model (Kriging surrogate) using samples

Fig. 1 Two steps of constructing a Kriging surrogate with five samples x = {0.10,0.30,0.56,0.74,0.98}. **a** Prior GP model with a quadratic trend function and 2σ intervals. **b** Kriging prediction and 2σ intervals after updating GP model

2.2.2 Simple discrepancy MFS framework with Kriging surrogate

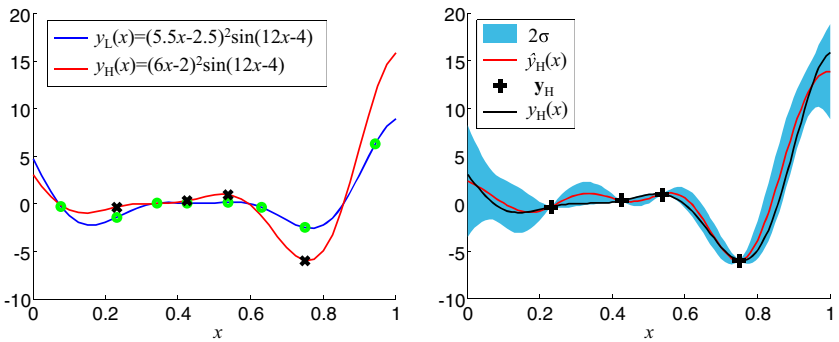
The combination of the simple framework and Kriging surrogates provides a special case of the GP based Bayesian discrepancy framework. This is because Kriging surrogates can be interpreted as single fidelity surrogates constructed with a GP based Bayesian framework as shown in Section 2.1. Since the low fidelity Kriging surrogate is constructed with a low fidelity sample set and the discrepancy Kriging surrogate is constructed with discrepancy between a low and high fidelity sample sets at common points, by using (1) (4) can be rewritten as

$$\hat{y}_H(\mathbf{x}) = \rho E\left(Y_L(\mathbf{x})\middle|\mathbf{y}_L\right) + E\left(\Delta(\mathbf{x})\middle|\boldsymbol{\delta}\right) \tag{5}$$

where $\mathbf{y}_L$ is a low fidelity data set; $\boldsymbol{\delta} = \mathbf{y}_H - \rho\mathbf{y}_L^c$ is a discrepancy sample set; $\mathbf{y}_L^c$ is a low fidelity data set at locations common to those of high fidelity data points; $\mathbf{y}_H$ is a high fidelity data set; and ρ is a regression scalar. We assume that the high fidelity model evaluated in a subset of the low fidelity evaluation points. Figure 2a shows low and high fidelity samples and the discrepancy sample set of the example has four elements obtained from the differences between high fidelity samples, and second, fourth, fifth and seventh low fidelity samples, respectively. $Y_L(\mathbf{x})$ and $\Delta(\mathbf{x})$ are GP models for the low fidelity and discrepancy function, and $Y_L(\mathbf{x})|\mathbf{y}_L$ and $\Delta(\mathbf{x})|\boldsymbol{\delta}$ are the corresponding GP models updated based on the sample sets. Note that the two GP models are expressed in different symbols but their functional form may be the same. For example, the two GP models can have the same linear polynomial trend functions and Gaussian correlation functions.

Figure 2a shows high and low fidelity samples with the corresponding functions. Figure 2b shows the prediction and the uncertainty of two standard deviations based on the samples. The variance of the total prediction uncertainty from the two surrogates is the sum of the two variances (that is the sum of the squares of the two standard errors). That is, the prediction uncertainty is measured by calculating square root of sums of squared standard errors of low fidelity and discrepancy Kriging surrogates. The root-mean-square-error (RMSE) was calculated using 1000 equally spaced points over [0,1]. Note that the high fidelity samples are a subset of the low fidelity sample set, from which discrepancy samples are obtained as a difference between them.

In the simple framework, the MFS is constructed with three steps: (1) the low fidelity Kriging surrogate is constructed with the low fidelity sample set, (2) the regression scalar ρ is determined by minimizing errors between the low fidelity surrogate and high fidelity sample set, and (3) a Kriging surrogate of the discrepancy function is



(a) A low and high fidelity functions and samples

(b) MFS fit (RMSE=0.89 for ρ=1.8)

Fig. 2 An MFS built with the simple framework. High and low fidelity DOEs are $\mathbf{X}_H$ = {0.54, 0.75, 0.43, 0.23} and $\mathbf{X}_L$ = {0.08, 0.23, 0.34, 0.43, 0.54, 0.63, 0.75 0.94}. **a** A low and high fidelity functions and samples. **b** MFS fit (RMSE = 0.89 for ρ = 1.8)

fitted using the difference between low and high fidelity sample set at the common sampling points. Note that the simple framework allows one to use any surrogate instead of Kriging, and even to use a different surrogate for the low fidelity fit and discrepancy fit.

2.2.3 Bayesian discrepancy MFS framework

The previously presented combination of the simple framework and Kriging surrogate can be generalized using GP based Bayesian approach, introduced by Kennedy and O'Hagan (2000) and Qian and Wu (2008). Note that this Bayesian framework provides equivalent predictions to the co-Kriging surrogate (Forrester et al. 2007; Le Gratiet 2013) with no prior information. This section gives emphasis to describing the difference between the GP based Bayesian discrepancy framework and the combination of the simple frameworks and Kriging.

The high fidelity GP model is defined as a linear combination of two GP models $Y_L(\mathbf{x})$ and $\Delta(\mathbf{x})$ as

$$Y_H(\mathbf{x}) = \rho Y_L(\mathbf{x}) + \Delta(\mathbf{x}) \tag{6}$$

where ρ is a regression scalar. Bayesian framework updates the GP model with the high and low fidelity sample sets. The mean of the updated GP model is used as predictor of the high fidelity response. The predictor is expressed as

$$\hat{y}_H(\mathbf{x}) = E\left(Y_H(\mathbf{x})\middle|\mathbf{y}_H, \mathbf{y}_L\right) \tag{7}$$

The GP models have the same form as in (2), which is composed of a trend function and a prediction uncertainty modeled with a GP random process. The GP models have the hyper parameters $\{\mathbf{b}_L, \sigma_L, \boldsymbol{\lambda}_L\}$ for the low fidelity model and $\{\mathbf{b}_\delta, \sigma_\delta, \boldsymbol{\lambda}_\delta\}$ for the discrepancy function model. The random processes of the two models are assumed to be independent (Kennedy and O'Hagan 2000).

When DOE samples are restricted such that a high fidelity sample set is a subset of a low fidelity sample set, the Bayesian discrepancy MFS becomes similar to the simple framework. It can be built by sequentially updating a low fidelity and discrepancy GP models. This sampling scheme also eases the computational burden of the full Bayesian approach since it allows sequential estimation of parameters. Hyper parameters of the low fidelity model are estimated first, and then, those of the high fidelity model and the regression scalar ρ are estimated. With the sampling restriction, (7) can be rewritten as

$$\hat{y}_H(\mathbf{x}) = \rho E\left(Y_L(\mathbf{x})\middle|\mathbf{y}_L\right) + E\left(\Delta(\mathbf{x})\middle|\boldsymbol{\delta}\right) \tag{8}$$

Note that the low fidelity model is updated by using only the low fidelity sample set since high fidelity data set has no effect on predicting the response of the low fidelity function. When the GP models of the Bayesian framework and those for Kriging surrogates of the simple framework are the same, the two frameworks make identical predictions about the low fidelity function.

The fitting process of the Bayesian discrepancy MFS takes two steps: (1) fitting the low fidelity GP model to low fidelity sample set by estimating the low fidelity hyper parameters and updating the GP model, and (2) fitting the discrepancy function GP model to the samples at the common points. With the sampling restriction, the MFS is the sum of these two updated models as (8) which is equivalent to the updated model of (6).

Note that Bayesian frameworks including this framework can incorporate a priori information for reducing the prediction uncertainty of an MFS through prior distributions for the hyper parameters. However, unlike physical parameters, where prior distributions can be obtained from experience, experience is rarely useful for selecting hyper parameters. Indeed, Co-Kriging, which is an MFS framework equivalent to the Bayesian discrepancy framework, has become popular even though it cannot incorporate prior distributions of hyper parameters. There is also an issue of objectivity in prior distributions in Bayesian estimation. In this paper, we use unbounded constant non-informative priors for hyper parameters for the Bayesian discrepancy MFS framework and the other Bayesian frameworks. For details about the hyper parameter estimation and the updating processes, readers are referred to Kennedy and O'Hagan (2000).

2.2.4 Difference between the simple and Bayesian discrepancy frameworks

It is difficult to see how the combination of the simple framework and Kriging surrogates and the Bayesian discrepancy framework would make different predictions with the same sample sets from (5) and (8). An important difference, which is not explicitly shown in the equations, is in determining the regression scalar, ρ. The simple framework determines ρ by maximizing the agreement between the scaled low fidelity fit and the high fidelity sample set. The Bayesian framework basically estimates ρ together with other hyper parameters of the discrepancy model as (7). Consequently, the Bayesian framework does not in general find ρ for maximizing the agreement.

To illustrate the difference between the two frameworks the high fidelity function used in Fig. 2 is employed again. The low fidelity function is a half of the high fidelity function plus a linear term, which compensates for the large difference between the two, as shown in Fig. 3.

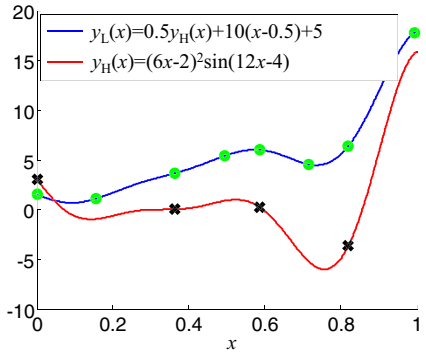


Fig. 3 High and low fidelity models to show the benefit of estimating ρ and $\hat{\delta}(x)$ simultaneously. High and low fidelity DOEs are $\mathbf{X}_H = \{0.00, 0.36, 0.59, 0.82\}$ and $\mathbf{X}_L = \{0.00, 0.15, 0.36, 0.49, 0.59, 0.71, 0.82, 0.99\}$

We use linear trend functions for the low and discrepancy GP models for the Bayesian framework and the Kriging surrogates for the simple framework. When ρ is estimated apart from the discrepancy, it needs to compensate for both the factor of two and the linear term. This results in a complicated discrepancy that cannot be approximated well with a small number of high fidelity samples. On the other hand, when both ρ and the discrepancy are estimated together, it is possible to take advantage of the simplicity of the true discrepancy.

Figure 4 shows the difference. Figure 4a and b present the two MFS predictions, Fig. 4c and d show the low fidelity function predictions from the Bayesian framework and the low fidelity Kriging surrogate for the simple framework, respectively. Figure 4e and f show the corresponding discrepancy predictions. The Bayesian framework yields an almost perfect fit, whereas the simple framework yields a bad prediction with an unrealistic prediction uncertainty estimation. As in (5) and (8), with the same GP models and DOE, the two frameworks provided identical predictions of the low fidelity function. The significant difference comes from the discrepancy function predictions. The simple framework found the regression scalar of 0.5 to maximize the agreement of the scaled low fidelity fit to the high fidelity samples, while the Bayesian framework found 2 that gives a simple linear discrepancy function. As shown in the low fidelity function formula in Fig. 3, the discrepancy function becomes a linear function by multiplying 2 on the low fidelity function. Therefore, the two frameworks made different estimations for the regression scalar and the discrepancy predictions.

This example possibly has an important implication for the simple framework with other surrogates. The essence is to find a ρ that will make the discrepancy y_H- ρy_L favorable to the discrepancy surrogate. For the simple framework with Kriging surrogate model, finding such a ρ could maximize the accuracy of the discrepancy surrogate. For that purpose, we may use a formulation that finds ρ minimizing RMSE of the trend function for the discrepancy y_H- ρy_L. With this formulation, the simple framework would find $\rho = 2$, which leads to a linear discrepancy function that can be fitted exactly with the linear trend function of the discrepancy Kriging surrogate. With this approach, an equivalent surrogate to the surrogate from the Bayesian framework can be built with the simple framework.

2.3 Calibration based frameworks

Calibration has been widely used to improve the predictability of a simulation by tuning physical parameters for the best agreement with samples from experiments (Kosonen and Shemeikka 1997; Owen et al. 1998; Lee et al. 2008; Coppe et al. 2012; Yoo and Choi 2013). MFS uses calibration to improve the prediction of a low fidelity simulation with samples from a high fidelity simulation (Ellis and Mathews 2001; Zheng et al. 2013). As with the previous discrepancy MFS frameworks, MFSs can be built with calibration.

In this paper, we first present a simple MFS framework using a surrogate-based calibration, followed by the Bayesian calibration framework based on a GP model. As with the previous discrepancy frameworks, we use a Kriging surrogate for the simple framework.

2.3.1 Simple calibration MFS framework

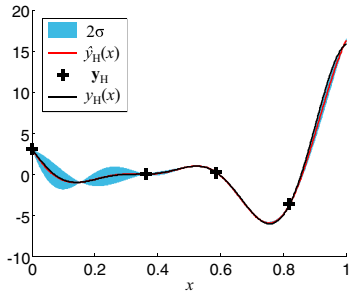
The simple calibration framework is to construct an MFS using a surrogate for a low fidelity model and to find the best model parameters for maximizing the agreement with high fidelity sample set. The calibrated high fidelity response is expressed as

$$\hat{y}_H(\mathbf{x}) = \rho \hat{y}_L(\mathbf{x}, \boldsymbol{\theta}) \tag{9}$$

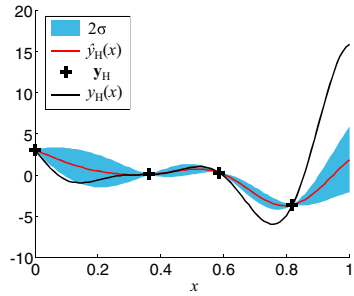
where Kriging surrogate is used for $\hat{y}_L(\mathbf{x}, \mathbf{q})$. We use a symbol $\mathbf{q}$ to denote variables for calibration parameters and $\boldsymbol{\theta}$ for their values when calibrated. The calibrated parameter vector $\boldsymbol{\theta}$ is obtained by minimizing the sum of squared discrepancies as

$$\boldsymbol{\theta} = \underset{\mathbf{q}}{\operatorname{argmin}} \sum_{i=1}^{n_H} \left(E\left(\rho Y_L(\mathbf{x}_h^i, \mathbf{q}) \middle| \mathbf{y}_L \right) - \mathbf{y}_h^i \right)^2 \tag{10}$$

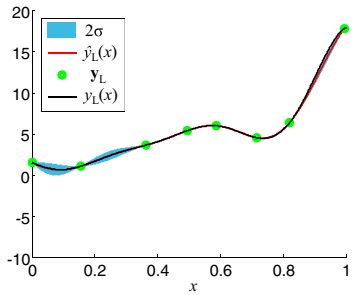
where $\mathbf{x}_h^i$ is the location of the ith high fidelity sample and n_H is the number of high fidelity samples. In this paper, a Kriging surrogate is used to construct a low fidelity surrogate for the simple calibration MFS framework, and the best calibration parameters can be obtained based on (10). Finally an MFS is obtained by substituting the parameters into (9).



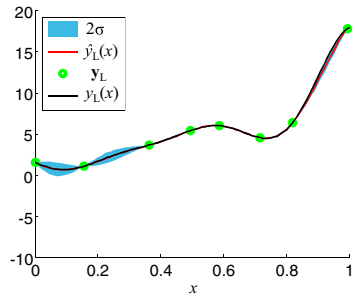
(a) $\hat{y}_H(x)$ and 2σ based on the Bayesian framework (RMSE=0.3 for ρ=2.0)



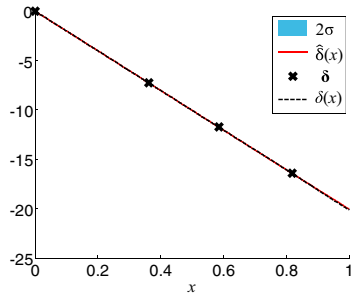
(b) $\hat{y}_H(x)$ and 2σ based on the simple framework with Kriging surrogate (RMSE=4.2 for ρ=0.5)



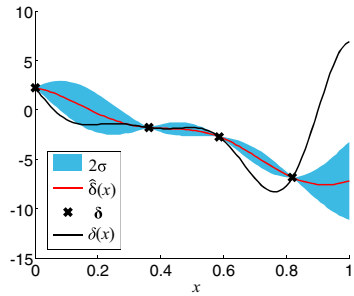
(c) $\hat{y}_L(x)$ and 2σ based on the Bayesian framework



(d) $\hat{y}_L(x)$ and 2σ based on the simple framework with Kriging surrogate



(e) $\hat{\delta}(x)$ and 2σ based on the Bayesian framework



(f) $\hat{\delta}(x)$ and 2σ based on the simple framework with Kriging surrogate

Fig. 4 An example of a substantial difference between the Bayesian and simple frameworks: The Bayesian discrepancy framework significantly outperforms the simple framework by finding ρ that makes the discrepancy more suitable to be fitted (as shown in (e)) rather than initially finding ρ by minimizing discrepancy between high fidelity samples and low fidelity function as shown in (f). **a** $\hat{y}_H(x)$ and 2σ based on the Bayesian framework (RMSE = 0.3 for $\rho = 2.0$). **b** $\hat{y}_H(x)$ and 2σ based on the simple framework (RMSE = 4.2 for $\rho = 0.5$). **c** $\hat{y}_L(x)$ and 2σ based on the Bayesian framework. **d** $\hat{y}_L(x)$ and 2σ based on the simple framework. **e** $\hat{\delta}(x)$ and 2σ based on the Bayesian framework. **f** $\hat{\delta}(x)$ and 2σ based on the simple framework

2.3.2 Bayesian calibration MFS framework

The Bayesian MFS framework makes inference about $y_H(\mathbf{x})$ based on both low and high fidelity sample sets (McFarland et al. 2008; Prudencio and Schulz 2012). A prediction is made by obtaining the mean of updated high fidelity GP model as in (7). The high fidelity GP model is defined using the low fidelity model and a regression scalar ρ as

$$Y_H(\mathbf{x}) = \rho Y_L(\mathbf{x}, \boldsymbol{\theta}) \tag{11}$$

As in the previous discrepancy Bayesian framework, the high fidelity GP model needs to be fitted to the samples for best consistency. To include the effect of calibration into the model, the low fidelity GP model is a function of **x** and **q** as the expression of Kriging surrogate for the previous simple framework. The low fidelity model is the sum of a trend function and a normal distribution $Z_L(\mathbf{x}, \mathbf{q}) \sim N(0, \sigma_L^2)$. The correlation function of $Z_L(\mathbf{x}, \mathbf{q})$ with n dimension of **x** and m dimension of **q** is defined as the product of two separate correlation functions.

$$\begin{aligned}&\operatorname{cov}\left(Z_L(\mathbf{x}, \mathbf{q}), Z_L\left(\mathbf{x}', \mathbf{q}'\right)\right) \\ &\quad = \sigma_L^2 \exp\left(-\left(\mathbf{x}-\mathbf{x}'\right)^T \boldsymbol{\Omega}_{\mathbf{x}} \left(\mathbf{x}-\mathbf{x}'\right)\right) \exp\left(-\left(\mathbf{q}-\mathbf{q}'\right)^T \boldsymbol{\Omega}_{\mathbf{q}} \left(\mathbf{q}-\mathbf{q}'\right)\right)\end{aligned} \tag{12}$$

where $\boldsymbol{\Omega}_{\mathbf{x}}$ and $\boldsymbol{\Omega}_{\mathbf{q}}$ are diagonal matrices. $\boldsymbol{\Omega}_{\mathbf{x}} = \operatorname{diag}(\lambda_{\mathbf{x},1}, \ldots, \lambda_{\mathbf{x},n})$ and $\boldsymbol{\Omega}_{\mathbf{q}} = \operatorname{diag}(\lambda_{\mathbf{q},1}, \ldots, \lambda_{\mathbf{q},m})$ denote diagonal matrices with n and m elements, respectively.

For details about the hyper parameter estimation process, readers are referred to McFarland et al. (2008) and Prudencio and Schulz (2012).

2.4 Comprehensive frameworks

The most flexible and widely used Bayesian comprehensive framework is the Bayesian calibration framework with a discrepancy function model, which is also known as Kennedy and O'Hagan framework (Kennedy and O'Hagan 2001a, b). The framework uses a GP model with Bayesian approach by considering the cross-interaction between the low fidelity and discrepancy function models. As a counterpart, we also describe a simple framework which builds an MFS using Kriging surrogates while ignoring the cross-interaction. To distinguish the two, we refer them as a simple comprehensive framework and a Bayesian comprehensive framework.

2.4.1 Simple comprehensive MFS framework

This framework applies the simple calibration framework first and fits the remaining difference with the simple discrepancy function framework with Kriging surrogates as

$$\hat{y}_H(\mathbf{x}) = \rho\hat{y}_L(\mathbf{x}, \boldsymbol{\theta}) + \hat{\delta}(\mathbf{x}) \tag{13}$$

where ρ and $\hat{y}_L(\mathbf{x}, \boldsymbol{\theta})$ are obtained as in the simple calibration framework and $\hat{\delta}(\mathbf{x})$ is constructed with a Kriging surrogate using the discrepancy between the high fidelity sample set and $\rho\hat{y}_L(\mathbf{x}, \boldsymbol{\theta})$.

2.4.2 Bayesian comprehensive MFS framework

The prediction of the Bayesian MFS is the mean of the posterior distribution obtained by updating the high fidelity model as in (7). The high fidelity model $y_H(\mathbf{x})$ is composed of the discrepancy function model and the low fidelity model. These two models are assumed to be independent. The high fidelity GP model is expressed as

$$Y_H(\mathbf{x}) = \rho Y_L(\mathbf{x}, \boldsymbol{\theta}) + \Delta(\mathbf{x}) \tag{14}$$

This comprehensive framework provides great flexibility, while it has the largest number of hyper parameters and model parameters. Therefore, the weakness of this framework is to estimate such many parameters simultaneously. Since it can be impractical to estimate all parameters simultaneously, it is possible to estimate them in groups (Kennedy and O'Hagan 2001a; Bayarri et al. 2007). For details about the estimation processes and the posterior distribution, readers are referred to Kennedy and O'Hagan (2001a, b).

2.5 Comparison between MFS frameworks using model calibration

In this section, we compare four MFS frameworks (two calibration and two comprehensive frameworks) via the example presented in the discrepancy function based framework section. The low and high fidelity functions defined in Fig. 3 were used in this section. The constants of the low fidelity function were replaced with calibration parameters and the low fidelity function is expressed as $y_{LT}(x, \theta_1, \theta_2) = 0.5y_{HT}(x) + \theta_1(x-0.5) + \theta_2$. Thus $\rho = 2$, $\theta_1 = 0$ and $\theta_2 = 0$ yield the exact high fidelity function for calibration based MFS frameworks with discrepancy function. The bounds of [0,10], [−5,5] and [1.5, 2.5] for θ_1, θ_2 and ρ were used, respectively.

A constant trend function was used for the low fidelity GP models and Kriging surrogates for the simple frameworks, and a linear polynomial trend function was used for discrepancy function GP models with the Gaussian correlation function. We generated 30 low fidelity samples in $\{x, \theta_1, \theta_2\}$ space using Latin hypercube sampling (LHS) as shown in Fig. 5. Four high fidelity samples were independently generated by LHS and then low fidelity sample evaluation points were updated using the nearest neighbor sampling. This process was repeated 100 times to generate 100 different low and high fidelity sets. The MFS frameworks were analyzed for each set of points. The differences between the simple and Bayesian frameworks were minor for most of the 100 sets, but for some sets they were substantial. Here, one case that shows the most notable difference is presented with a high fidelity sample set for $\mathbf{x}_H = \{0.06, 0.35, 0.69, 0.9\}^T$. Figure 6 compares the performance of four frameworks with the identified values of calibration parameters.

Figure 6a shows the simple calibration MFS. Even if the identified model parameters (θs) are very close to the exact values, the fitted MFS is not correspondingly close to the true high fidelity function due to a finite number of samples used to construct the low fidelity surrogate. The 2σ shaded region is twice the standard error of the predicted high fidelity response, which is obtained by multiplying ρ and standard error of the low fidelity Kriging surrogate based on (9). Note, however, that the uncertainties in the calibrated parameters are not taken into account in this shading.

A notable difference between the simple and Bayesian calibration frameworks is that the Bayesian frameworks make the prediction to pass exactly through the high fidelity samples

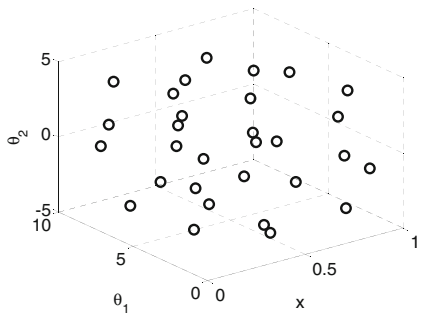
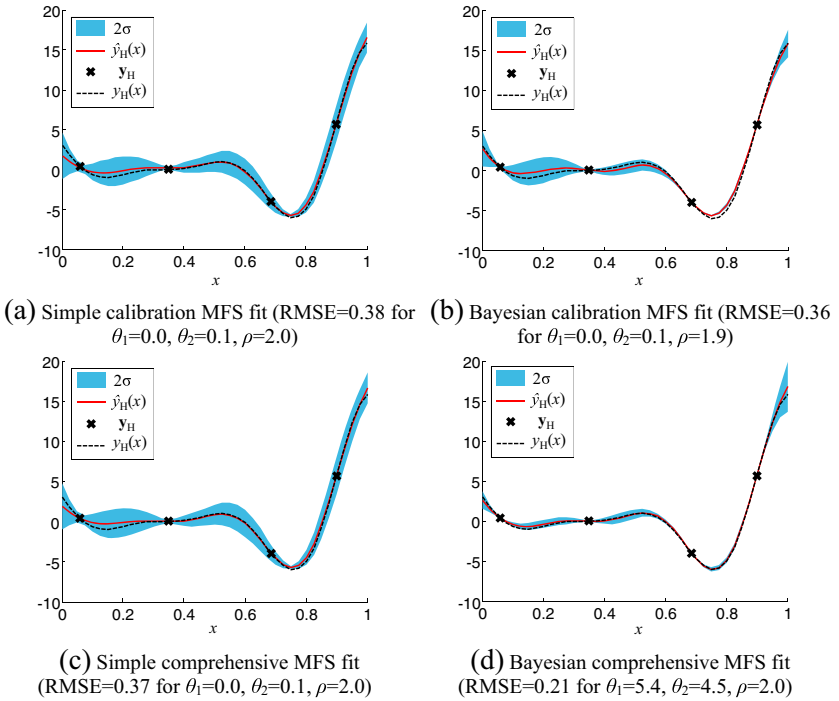


Fig. 5 Thirty low fidelity samples in $\{x, \theta_1, \theta_2\}$ space generated by Latin Hypercube Sampling



(a) Simple calibration MFS fit (RMSE=0.38 for θ_1=0.0, θ_2=0.1, ρ=2.0)

(b) Bayesian calibration MFS fit (RMSE=0.36 for θ_1=0.0, θ_2=0.1, ρ=1.9)

(c) Simple comprehensive MFS fit (RMSE=0.37 for θ_1=0.0, θ_2=0.1, ρ=2.0)

(d) Bayesian comprehensive MFS fit (RMSE=0.21 for θ_1=5.4, θ_2=4.5, ρ=2.0)

Fig. 6 Multi-fidelity surrogates with model calibration: This example shows that the Bayesian comprehensive MFS framework finds parameters different from the other frameworks. **a** Simple calibration MFS fit (RMSE = 0.38 for $\theta_1 = 0.0$, $\theta_2 = 0.1$, $\rho = 2.0$). **b** Bayesian calibration MFS fit (RMSE = 0.36 for $\theta_1 = 0.0$, $\theta_2 = 0.1$, $\rho = 1.9$). **c** Simple comprehensive MFS fit (RMSE = 0.37 for $\theta_1 = 0.0$, $\theta_2 = 0.1$, $\rho = 2.0$). **d** Bayesian comprehensive MFS fit (RMSE = 0.21 for $\theta_1 = 5.4$, $\theta_2 = 4.5$, $\rho = 2.0$)

while the simple frameworks do not. In Fig. 6a and c, the predictions do not pass the high fidelity samples exactly. Thought the difference is not visible in the figure. The characteristics are reflected in their prediction variances that the prediction variances of the Bayesian frameworks are zero values at high fidelity samples but those of the simple frameworks are nonzero values.

The Bayesian comprehensive framework has a distinct characteristics of tuning parameters compared to other frameworks. In terms of RMSE, the Bayesian comprehensive framework significantly outperforms other frameworks. However, the Bayesian comprehensive framework drew very different parameters from the exact ones, while other frameworks came close to the exact values. This is because of the presence of the discrepancy function model in the framework. When $\rho = 2$, the discrepancy function becomes a linear polynomial. Since the Bayesian comprehensive framework can perfectly fit the linear discrepancy, it tunes parameters in order to give a best fit regardless the tuned parameters maximize the agreement of the low fidelity model or not, while other frameworks tune parameters to maximize the agreement.

2.6 Design of experiments for MFS surrogate frameworks

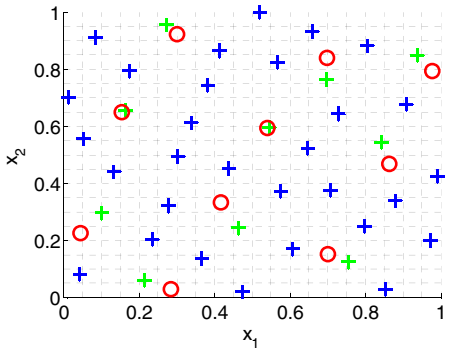
For simple discrepancy function based frameworks, the sampling condition that the low fidelity sample set is a super set of the high fidelity sample set is common to observe model discrepancy at common points (Balabanov et al. 1998 and Mason et al. 1998). The same assumption is favored by the Bayesian framework (Kennedy and O'Hagan 2000).

The nearest neighbor sampling takes two steps. Firstly it generates independent initial low and high fidelity sampling points using LHS. In the second step move each low-fidelity point to the nearest high fidelity sampling point (Le Gratiet 2013). Figure 7 shows an example of 40 low fidelity and 20 high fidelity sampling points based on the nearest neighbor design.

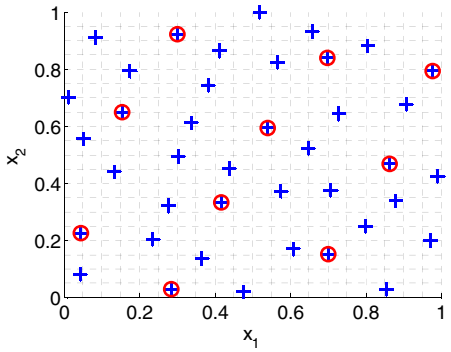
Another sampling strategy for achieving the sampling condition is the nested design sampling, which was initially developed as a space filling technique for adding samples to an existing sample set as a way to ensure optimal coverage of the union of two sample sets (Jin et al. 2005). This strategy can be applied to generate low and high fidelity sampling points. Sampling points for fitting a low fidelity surrogate is generated using LHS and additional sampling points are generated by maximizing the minimum distance between all existing and new points. The union of the existing sampling point set and the additional sampling points is the low fidelity sampling point set and the additional sampling point set is the high fidelity sample set.

These two sampling strategies provide low and high fidelity sample sets that satisfy the sampling condition of the high-fidelity sampling points being a subset of the low-fidelity sampling points. In this paper, the nearest neighbor design is used for the discrepancy function based frameworks for the numerical examples.

For calibration frameworks, the low-fidelity samples include both input variables and calibration parameters, while the high-fidelity samples need to select only input variables. We generated low fidelity sampling points using LHS in the



(a) 40 low fidelity (blue and green crosses) and 10 high fidelity (red circles) sampling points; the green crosses (nearest neighbors of the high fidelity sampling points) are replaced with the high fidelity sampling points



(b) Updated 40 low fidelity sampling points (blue crosses) and 10 high fidelity sampling points (red circles) obtained using the nearest neighbor sampling technique

Fig. 7 The nearest neighbor sampling for the discrepancy function based frameworks. **a** 40 low fidelity (*blue and green crosses*) and 10 high fidelity (*red circles*) samples; the green crosses (nearest neighbors of the high fidelity sampling points) are replaced with the high fidelity samples. **b** Updated 40 low fidelity samples (*blue crosses*) and 10 high fidelity samples (*red circles*) obtained using the nearest neighbor sampling technique

dimensions of the input variables and the calibration parameters. Then high fidelity sampling points were generated in the dimensions of the input variables using LHS. The coordinates of the low fidelity sampling points in the dimension of the input variables were adjusted using the high fidelity sampling points and the nearest neighbor sampling technique. Figure 8 shows the low and high fidelity sampling points for 2 input variables and 1 model parameter for a calibration based MFS framework. Figure 8a shows 40 low fidelity samples obtained from LHS and Fig. 8b shows the 40 samples projected on the input dimensions and 20 high fidelity sampling points obtained by the nearest neighbor sampling.

3 Performance measures for MFS frameworks

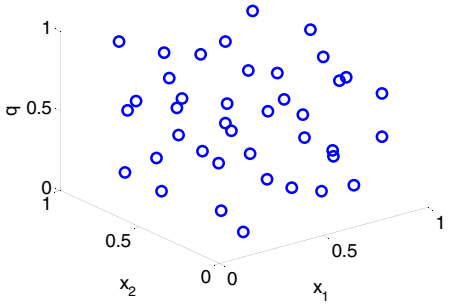
In this section, we describe performance measures for comparing MFS frameworks. Since the performance of surrogates varies for different design of experiments (DOE), we collect statistics of frameworks for 100 different DOEs randomly generated using the nearest neighbor design. We evaluate the accuracy of an MFS model for given cost and the cost for given accuracy as well as the variability in these measures. Since the cross validation error has been considered as a useful performance measure for surrogates (Sanchez et al. 2008; Acar and Rais-Rohani 2009; Viana et al. 2009), we investigated whether it can also be employed for MFSs to rank the frameworks. We ranked frameworks for a given DOE using their cross validation errors and compared the rank to the actual ranks based on their RMSEs, which is our reference accuracy metric.

3.1 Measures of accuracy

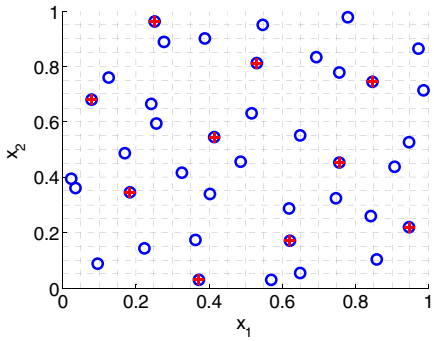
MFSs are fitted to function values at n sample points. We can measure the accuracy of an MFS using the root mean square error (RMSE) integrated over the sampling domain. We use Monte Carlo integration with n_{test} test points as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(y^i_{h,test} - \hat{y}_H \left(\mathbf{x}^i_{h,test} \right) \right)^2} \quad \text{for } i = 1, \ldots n_{test} \qquad (15)$$

where $\mathbf{x}^i_{h,test}$ and $y^i_{h,test}$ are the sampling point vector and the high fidelity function value of i^{th} test point. Note that RMSE is only possible to calculate when the true function is available.



(a) 40 low fidelity samples for 2 input variables (x_1 and x_2) and 1 calibration parameter (q)



(b) Projected 40 low fidelity samples (blue circles) and 10 high fidelity samples (red crosses) on the x_1-x_2 plane

Fig. 8 Low fidelity samples (*blue*) and high fidelity samples (*red*) using nearest neighbor sampling used for the calibration based frameworks. **a** 40 low fidelity samples for 2 input variables (x_1 and x_2) and 1 calibration parameter (q). **b** Projected 40 low fidelity samples (*blue circles*) and 10 high fidelity samples (*red crosses*) on the x_1-x_2 plane

3.2 Cross validation errors

Among all available frameworks, how to select the best framework or how to filter out bad frameworks is an important question. Since the cross validation error performs well for selecting surrogates (Viana et al. 2009), it is also considered here for the MFS frameworks. For single fidelity surrogates, a leave-one-out cross validation error is the error measured at a sample point using the surrogate constructed with all samples except for the point. By repeating this process for every sample, we can estimate n cross validation errors, and the RMSE of a surrogate can be estimated by calculating the global cross-validation error measure called *PRESS* using cross validation errors (Viana et al. 2009).

Since the discrepancy MFS frameworks use samples with the sampling restriction that a high fidelity sampling point set is a subset of a low fidelity sampling set, a cross validation error of a discrepancy MFS is obtained by leaving out both low and high fidelity samples at a common point (Le Gratiet 2013). For the calibration based frameworks, since the low fidelity sample set has dimensions for input variables and calibration parameters while the high fidelity sample set has dimensions for only input variables. For example, Fig. 8 shows a low sample set for input variables x_1 and x_2, and calibration parameter q and a high fidelity sample set for input variables. We leave out only high fidelity samples for obtaining cross validation errors. Therefore, the number of cross validation errors is same as that of high fidelity samples. The RMSE of an MFS framework is estimated by calculating *PRESS* as

$$PRESS = \sqrt{\frac{1}{n_H}\sum_{i=1}^{n_H} e_i^2} \tag{16}$$

where e_i is the cross validation error by leaving out the i^{th} high fidelity sample.

Instead of omitting one sample point at a time, it is possible to exclude a group of samples at a time. k-fold cross validation error randomly divides the sample set into k subsets and one of the subsets is left out to calculate a cross validation error. For k-fold cross validation, $\mathrm{n_H/k}$ cross validation errors in (16) are simultaneously obtained from k repetitions instead of n_H repetitions for the one leaving out strategy. For example, with 10 high fidelity samples divided into 5 sets as $\{(\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_3, \mathbf{x}_4), \ldots, (\mathbf{x}_9, \mathbf{x}_{10})\}$, cross validation errors e_1 and e_2 are obtained by calculating the errors against the surrogate constructed with remaining four groups of samples. This process is repeated five times to obtain all the cross validation errors and *PRESS*. We implement a strategy to divide samples into k-folds using a "maximin" criterion (maximization of the minimum inter-distance) (Viana et al. 2009), which is used in the following numerical examples. Note that we use unbounded constant non-informative priors for the Bayesian frameworks. Therefore the influence of error in a priori information is excluded in the constructed MFSs and the same for *PRESS*.

4 Numerical examples

In this section, we statistically evaluate the performances of frameworks based on randomly generated DOEs. The Hartmann 6 function and the borehole function were used for the statistical study of frameworks. The statistical study of the previously presented 1D example is also presented in Appendix A. Hartmann 6 function is an algebraic function with 6 input variables. The Borehole function is a physical function with 3 input variables initially developed to calculate the flow of water through a borehole drilled from the ground surface through two aquifers.

We refer the frameworks and their variants with labels, which are presented in Table 1. First letters "S" and "B" indicate that a framework uses simple framework using ready-made surrogates or Bayesian framework, respectively. Then correction approaches are indicated by "C" and "D" for calibration and discrepancy function, respectively. For the comprehensive frameworks, which use both approaches, we use the letters together as "CD". To test the effect of including a regression scalar ρ in a framework, we compared MFSs constructed with and without ρ. We remove the scaling effect of using a regression scalar by applying $\rho = 1$ to a framework. The last letter "R" indicates that ρ is included in a framework.

Table 1 Acronyms of frameworks used

Framework	Label
Low fidelity surrogate	L
High fidelity surrogate	H
Simple discrepancy	SDR
	SD ($\rho = 1$)
Bayesian discrepancy	BDR
	BD ($\rho = 1$)
Simple calibration	SCR
	SC ($\rho = 1$)
Bayesian calibration	BCR
	BC ($\rho = 1$)
Simple comprehensive	SCDR
Bayesian comprehensive	BCDR

4.1 Hartmann 6 function

MFS frameworks are often employed to reduce computational cost and/or to improve accuracy. In this section, we examine MFS frameworks with three factors that affect the performance: 1) total computational budget, 2) high-to-low sample evaluation cost ratio, and 3) high-to-low sample size ratio.

Table 2 shows combinations of low and high fidelity samples for given total budget and cost ratio. Since the key question in MFS is whether low fidelity simulations would reduce the cost of high fidelity simulations, the total computational budget is expressed in terms of the number of high fidelity samples. For example, 56H denotes that we have computational budget for evaluating 56 high fidelity samples. Sample cost ratio tells how many low fidelity samples can be evaluated for the cost for a single high fidelity sample. For example, cost ratio of 4 means that 4 low fidelity samples can be evaluated with the budget for evaluating a single high fidelity sample. For that ratio, a total budget of 56H can be used for either 56 high fidelity samples, or 224 low fidelity samples, or mixes of the two such as the 36 high-fidelity and 80 low fidelity samples shown in Table 2. These combinations are expressed with the numbers of high (n_H) and low (n_L) fidelity samples, such as 36/80. We selected mixes ranging from spending most of the budget on high-fidelity simulations (36/80), to spending most of it on low fidelity simulation (6/200).

We use the Hartmann 6 function over [0.1,1] for all dimensions as a high fidelity function and an approximated function as a low fidelity function. As we did for the previous example, for each case, we generated 100 different DOEs using the nearest neighbor sampling and LHS. RMSE was calculated based on 10,000 test points which give an accurate RMSE estimate. Among 100 RMSEs, the median RMSE was chosen as a representative value. A median RMSE is the RMSE greater than lower 50 % RMSE population. Also median is less sensitive to extreme values than mean. For the comparison's sake, the same test points were used to calculate the RMSEs for different DOEs.

Table 2 Cases of sample cost and size ratios combinations for two total computational budgets (Hartmann 6 function example)

Total budget	Sample cost ratio	Sample size ratio n_H/n_L
56H	4	36/80, 26/120, 16/160, 6/200
	10	49/70, 46/100, 42/140, 35/210, 28/280, 21/350, 14/420, 7/490
	30	48/240, 46/300, 44/360, 42/420, 40/480, 38/540, 28/840, 18/1140
28H	4	18/40, 13/60, 8/80, 3/100
	10	22/60, 19/90, 16/120, 13/150, 10/180, 7/210, 4/240
	30	24/120, 22/180, 20/240, 18/300, 16/360, 14/420, 10/540, 6/660, 4/720

The high fidelity function of this example, Hartmann 6 function, is

$$f_H(\mathbf{x}) = -\frac{1}{1.94}\left(2.58 + \sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{6} A_{ij}\left(x_j - P_{ij}\right)^2\right)\right) \tag{17}$$

where $\boldsymbol{\alpha} = \{1 \quad 1.2 \quad 3 \quad 3.2\}^T$ are model parameters, and the following A and P matrices are constant:

$$\mathbf{A} = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix} \text{ and } \mathbf{P}$$

$$= 10^{-4}\begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix} \tag{18}$$

The low fidelity function is

$$f_L(\mathbf{x}) = -\frac{1}{1.94}\left(2.58 + \sum_{i=1}^{4} \alpha_i' f_{\exp}(\nu_i)\right) \text{ and } \nu_i$$

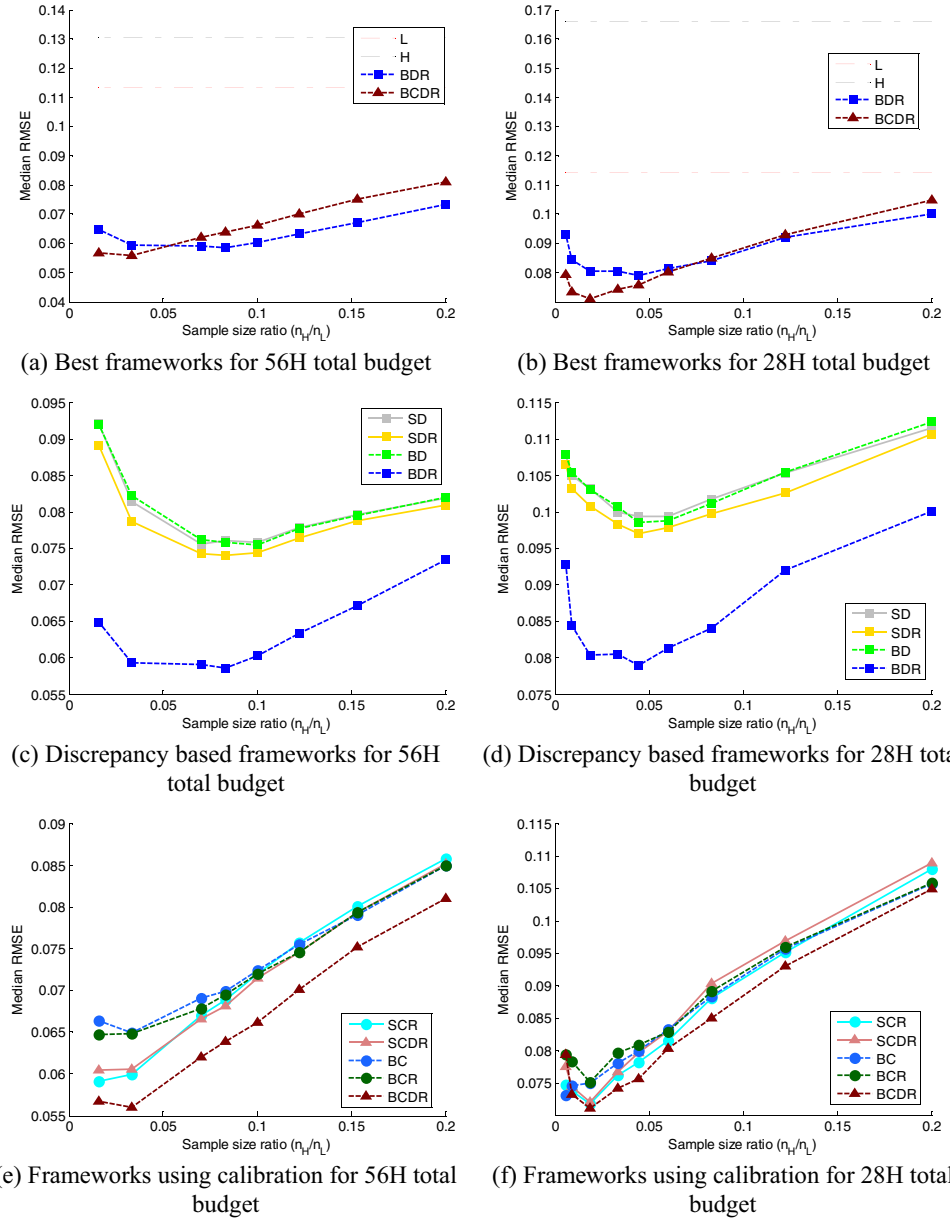
$$= -\sum_{j=1}^{6} A_{ij}\left(x_j - P_{ij}\right)^2 \tag{19}$$

where $\boldsymbol{\alpha}' = \{0.5 \quad 0.5 \quad 2.0 \quad 4.0\}^T$ and $f_{\exp}(x)$ is the approximation function of the exponential function which is expressed as

$$f_{\exp}(x) = \left(\exp\left(\frac{-4}{9}\right) + \exp\left(\frac{-4}{9}\right)\frac{(x+4)}{9}\right)^9 \tag{20}$$

Note that the total function variation of the Hartmann 6 function is 0.33 and the RMSE of the low fidelity function with respect to the high fidelity function is 0.11. That means the maximum RMSE of a surrogate based solely on low fidelity samples is 0.11. The RMSE was obtained using 10,000 test points.

Bounds of [0.5, 1.5] were used for SDR, BDR, SCR, BCR, SCDR and BCDR, which use a regression scalar. For the frameworks using calibration, α_3' and α_4' were selected as model parameters to be calibrated. The bounds of [1, 6] and [2, 7] were used for the parameters. Note that we use 0.5 and 0.5 for α_1' and α_2', respectively. We only calibrate α_3' and α_4' to simulate a common situation that we cannot afford to calibrate all parameters.



(a) Best frameworks for 56H total budget

(b) Best frameworks for 28H total budget

(c) Discrepancy based frameworks for 56H total budget

(d) Discrepancy based frameworks for 28H total budget

(e) Frameworks using calibration for 56H total budget

(f) Frameworks using calibration for 28H total budget

Fig. 9 Median RMSEs for different sample size ratios and cost ratio of 30. **a** Best frameworks for 56H total budget. **b** Best frameworks for 28H total budget. **c** Discrepancy based frameworks for 56H total budget. **d** Discrepancy based frameworks for 28H total budget. **e** Frameworks using calibration for 56H total budget. **f** Frameworks using calibration for 28H total budget

Figure 9 shows the effect of sample ratio on the median of RMSE using (15) for the cases of total computational budgets 56H and 28H with cost ratio of 30. A similar trend is observed for the cost ratio of 10, which is shown in Fig. 13 in Appendix B. Figure 9a and b compare the median RMSE of the best frameworks selected from the discrepancy function-based frameworks and the frameworks using calibration to that of the single fidelity surrogates. The red and black dashed lines represent the median RMSEs of the single high fidelity and low-fidelity Kriging surrogates, respectively. The median RMSEs were obtained from Kriging surrogates constructed with randomly generated 100 DOEs using LHS. For the single fidelity surrogates, total budget was used to achieve samples of one fidelity; there is no effect of sample-size ratio. With a cost ratio of 30, we can afford $56 \times 30 = 1680$ low fidelity simulations which give us a median RMSEs close to the true RMSE of 0.11 for exact low-fidelity function. This means, that for such high-cost ratio we can expect better accuracy from low fidelity surrogates than high fidelity surrogates. Figure 9c and d present the median RMSE of the discrepancy function based frameworks and Fig. 9e and f present that of the frameworks using calibration.

The overall observation about results with the total budget of 56H and 28H are similar. All the MFS frameworks were better in terms of median RMSEs than the Kriging surrogates using only low or high fidelity samples for most of sample size ratios. BDR framework performed best for mostly sample size ratios less than 0.1. The frameworks using calibration outperformed BDR for small sample ratios while the performance of BDR rapidly decreased as sample size ratio

decreased. Each framework had an optimal sample size ratio. For this example, the optimal sample size ratio of BDR was around 0.1 and those of the frameworks using calibration were much lower than that. Note that, for the sample size ratio 0, the median RMSE of each framework converged to that of low fidelity surrogates. Note that for a cost ratio of 30, most of the budget was still spent on high-fidelity simulations for sample size ratio larger than 0.33.

For this example, combination of full Bayesian approach and a regression scalar ρ improved accuracy of the discrepancy function-based frameworks significantly. For all cases, the performance of BDR was significantly better than that of BD, SD and SDR. That means applying Bayesian (BD to BDR) without ρ or the other way (SDR to BDR) did not make noticeable improvement but combination of Bayesian and ρ works for this example. BCDR worked generally well among the frameworks using calibration. For 28H and cost ratio of 10, BDR performs best for sample size ratio larger than 0.13 and BCDR performs best for sample size ratio less than 0.13.

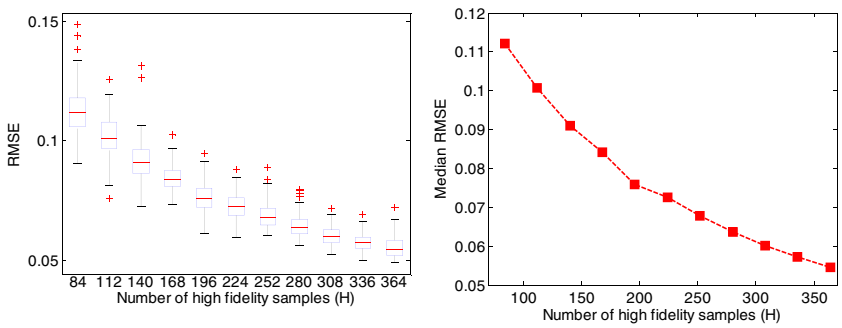
SCR outperformed BCR for small sample size ratios ($\leq$0.1), while the two behaved similarly for larger ratios. BCDR outperformed SCDR for almost all the diagramed sample size ratios. Unlike the discrepancy based frameworks, the effect of a regression scalar ρ was limited for the frameworks using calibrations. Unlike the discrepancy based frameworks, the effect of a regression scalar ρ was limited for the frameworks using calibrations.

Each framework had its break-even sample size ratio where the median RMSE of an MFS framework becomes smaller than the best low fidelity median RMSE. The break-even sample size ratio decreased as the total budget decreased. For 56H and cost ratio of 10, most of breaking even points were within the sample size ratios 0.4 and 0.8 while, for 28H and cost ratio of 10, breaking even points were within the sample size ratios 0.2 and 0.4. That means securing enough number of low fidelity samples was important for this example. Little benefit was expected by applying MFS frameworks for large sample size ratios.

In this paper, cost saving of an MFS framework is calculated by calculating computational cost to achieve the equivalent median RMSE. For example, the low fidelity median RMSE of BDR for the case of 56H and 30 cost ratio is 0.059. Since such low level of median RMSE cannot be achieved with the low fidelity surrogate, we calculated the number of high fidelity samples that provides an equivalent median RMSE.

Figure 10a shows boxplots of RMSEs of 100 Kriging surrogates built with 100 different high fidelity sample sets, which were randomly generated using LHS. Figure 10b shows the corresponding median RMSE graph in terms of the number of high fidelity samples. We obtain the cost saving using Fig. 10b. For example, the best median RMSE of BDR for the case of 56H and 30 cost ratio is 0.059 based on Fig. 9c. We calculate the number of high fidelity samples which gives 0.059 in the graph which is 320H. The BDR framework can achieve the equivalent median RMSE with 56H, so the cost saving is −83 %.

Table 3 shows the summary of the results for different cases. For each case, we selected the best frameworks in terms of median RMSE. Commonly, frameworks using calibration performed best for small sample size ratio, and BDR performed generally well. Best frameworks for different sample size ratio ranges are presented. For example, for the combination of total budget 56H and cost ratio of 4, BC is the best framework for the sample size ratio smaller than 0.1, and BDR is the best for larger than 0.1. The maximum median RMSE improvement and the cost saving in the corresponding ranges were also presented. The median RMSE of an MFS framework is a function of the sample size ratio. We found the minimum median RMSE of each framework over all considered sample size ratios. For each MFS framework, the difference between its minimum median RMSE and the median RMSE of Kriging surrogates using only high fidelity samples was calculated and considered as maximum median RMSE reduction. Cost savings are then calculated by finding how many high fidelity samples for Kriging surrogates are needed to achieve the same minimum median RMSE was calculated. Note that the median RMSEs of MFS frameworks were smaller than the RMSE of the low fidelity function that is the minimum achievable RMSE with only low fidelity samples.



(a) Box plots of RMSEs (b) Median RMSEs

Fig. 10 RMSE of Kriging fits built based on only high fidelity samples. **a** Box plots of RMSEs. **b** Median RMSEs

Table 3 Summary of the statistical studies (Hartmann 6 function example)

Total computational budget	28H					
Cost ratio	4		10		30	
Best single fidelity surrogate (median RMSE)	L (0.136)		L (0.121)		L (0.114)	
For sample size ratio	≥0.1	<0.1	≥0.13	<0.13	≥0.08	<0.08
Best frameworks (min. median RMSE)	BDR (0.125)	BC (0.12)	BDR (0.11)	BC (0.09)	BDR (0.08)	BCDR (0.075)
Max. median RMSE reduction in %	−8 % [a]	−12 %	−9 %	−26 %	−30 %	−34 %
Max. cost saving in % (number of HF samples for equivalent median RMSE)	−59 % [b] (68)	−62 % (74)	−69 % (90)	−81 % (144)	−85 % (182)	−86 % (203)
Total computational budget	56H					
Cost ratio	4		10		30	
Best single fidelity surrogate (median RMSE)	L (0.123)		L (0.115)		L (0.113)	
For sample size ratio	≥0.1	<0.1	≥0.1	<0.1	≥0.07	<0.07
Best frameworks (min. median RMSE)	BDR (0.095)	BC (0.1)	BDR (0.077)	BCDR (0.075)	BDR (0.059)	BCDR (0.055)
Max. median RMSE improvement in %	−23 %	−19 %	−33 %	−35 %	−48 %	−51 %
Max. cost saving in % (number of HF samples for equivalent median RMSE)	−56 % (128)	−51 % (114)	−71 % (192)	−72 % (203)	−83 % (320)	−85 % (364)

[a] The minimum median RMSE for the case of 28H and cost ratio of 4 is 0.125 and that of the best single fidelity surrogate is 0.136, so that maximum median RMSE reduction was calculated as (1–0.125/0.136)*100

[b] The number of high fidelity samples to achieve the same minimum RMSE is 68H, the corresponding cost saving is calculated as (1-28H/68H)*100

This example shows that the benefit of employing MFS frameworks for saving computational cost was higher than for improving the accuracy. MFS frameworks are mostly beneficial for high cost ratios. The observed maximum cost saving is −86 % for the case of 28H total budget and 30 cost ratio. The maximum RMSE improvement is −51 % and the corresponding best median RMSE is 0.055. The MFS frameworks are generally useful for high cost ratio cases.

4.2 Borehole function

In this section, we examine the MFS frameworks with a physical example, the flow of water through a borehole penetrating two aquifers. We considered the combinations of low and high fidelity samples for given total budget and cost ratio shown in Table 4. The same notation is used for total budget, cost ratio and sample size ratio as with the previous example.

The model for the flow was obtained based on assumptions of steady-state flow from the upper aquifer into the borehole and from the borehole into the lower aquifer, no groundwater gradient, and laminar, isothermal flow through the borehole (Morris et al. 1993). The flow rate through the borehole in m^3/yr for given conditions was obtained as

$$f_H(R_w, L, K_w) = \frac{2\pi T_u (H_u - H_l)}{\ln\left(R/R_w\right)\left(1 + \frac{2LT_u}{\ln\left(R/R_w\right)R_w^2 K_w} + \frac{T_u}{T_l}\right)} \quad (21)$$

We assume that the flow rate $f_H(R_w, L, K_w)$ is calculated for input variables R_w, L, and K_w defining the dimensions and conductivity of a borehole and the other environmental parameters were measured and given. The input variables and parameters of the function are presented in Table 5. The values of the parameters were set to the nominal values of interest ranges of the parameters. Morris et al. (1993) presents the interest ranges of the parameters.

Table 4 Cases of sample cost and size ratios combinations for three total computational budgets (Borehole function example)

Total budget	Sample cost ratio	Sample size ratio
5H	10	4/10, 3/20, 2/30
	30	4/30, 3/60, 2/90
10H	10	8/20, 7/30, 6/40, 5/50, 4/60, 3/70, 2/80
	30	9/30, 8/60, 7/90, 6/120, 5/150, 4/180, 3/210, 2/240

Table 5 Input variables and environmental parameters

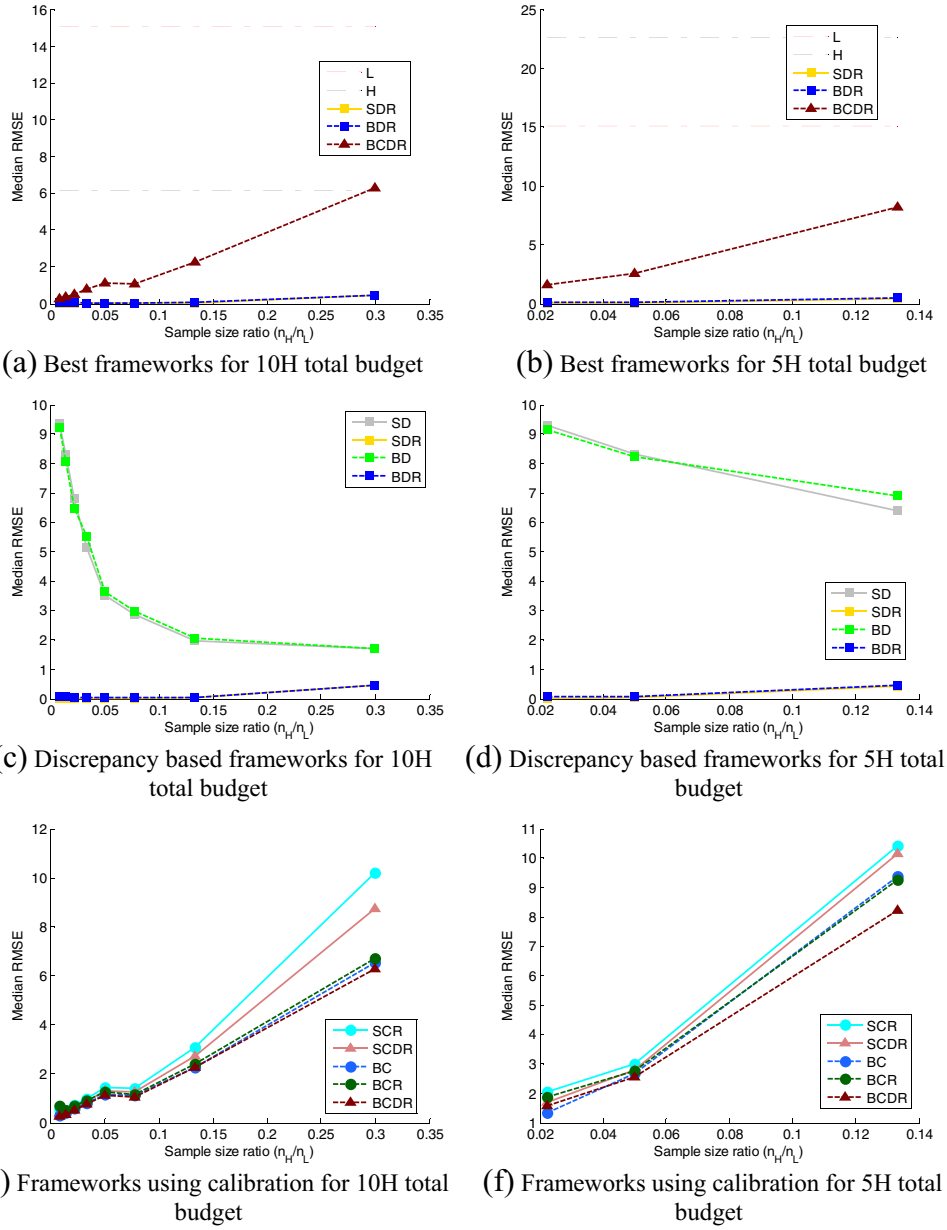
Input	Description
$R_w = [0.05, 0.15]$ m	Radius of borehole
$L = [1120,1680]$ m	Length of borehole
$K_w = [1500,15000]$ m/year	Hydraulic conductivity of borehole
Parameters	
$R = 25050$ m	Radius of influence
$T_u = 89335$ m^2/year	Transmissivity of upper aquifer
$H_u = 1050$ m	Potentiometric head of upper aquifer
$T_l = 89.55$ m^2/year	Transmissivity of lower aquifer
$H_l = 760$ m	Potentiometric head of lower aquifer

The borehole function was used as a high fidelity function and we selected a low fidelity function from a literature (Xiong et al. 2013). The RMSE of the low fidelity function with respect to the high fidelity function is 15.0. The low fidelity function is expressed as

$$f_L(R_w, L, K_w) = \frac{5T_u(H_u - H_l)}{\ln\left(R/R_w\right)\left(1.5 + \frac{2LT_u}{\ln\left(R/R_w\right)R_w^2 K_w} + \frac{T_u}{T_l}\right)} \quad (22)$$

For all frameworks, regression scalar (ρ) bounds of [0.5, 1.5] were used. H_u and H_l were selected as model parameters to be tuned since the flow rate is sensitive to them. [800, 1200] and [600, 1000] were used as the bounds of the parameters, H_u and H_l, respectively.

Figure 11 shows the effect of sample ratio on the median of RMSE for different cases. The cases of total computational budgets 5H and 10H with cost ratio of 30 are shown. Figure 11a and b show the median RMSE of the best frameworks, and the red and black dashed lines represent the median RMSEs of 100 single low and high fidelity Kriging surrogates, respectively. The median RMSE of low fidelity surrogates is almost the same with the RMSE of the low fidelity function of 15.0 for this case. That means the low fidelity surrogates have almost the same error as the exact low fidelity function. For 10H



(a) Best frameworks for 10H total budget

(b) Best frameworks for 5H total budget

(c) Discrepancy based frameworks for 10H total budget

(d) Discrepancy based frameworks for 5H total budget

(e) Frameworks using calibration for 10H total budget

(f) Frameworks using calibration for 5H total budget

Fig. 11 Median RMSEs for different sample size ratios and cost ratio of 30. **a** Best frameworks for 10H total budget. **b** Best frameworks for 5H total budget. **c** Discrepancy based frameworks for 10H total budget. **d** Discrepancy based frameworks for 5H total budget. **e** Frameworks using calibration for 10H total budget. **f** Frameworks using calibration for 5H total budget

Table 6 Best frameworks for different sample size ratios and the corresponding cost saving (Borehole function example)

Total computational budget	5H			
Cost ratio	10		30	
Best single fidelity surrogate (median RMSE)	L (15.0)		L (15.0)	
Best frameworks (min. median RMSE)	SDR (0.4)	BDR (0.5)	SDR (0.00)	BDR (0.08)
Max. median RMSE improvement in %	−97 %	−97 %	−100 %	−100 %
Max. cost saving in % (number of HF samples for equivalent median RMSE)	−84 % (32)	−83 % (30)	> − 95 % (>100)	−90 % (50)
Total computational budget	10H			
Cost ratio	10		30	
Best single fidelity surrogate (median RMSE)	H (6.1)		H (5.7)	
Best frameworks (min. median RMSE)	SDR (0.02)	BDR (0.07)	SDR (0.00)	BDR (0.05)
Max. median RMSE improvement in %	−100 %	−99 %	−100 %	−99 %
Max. cost saving in % (number of HF samples for equivalent median RMSE)	−85 % (70)	−80 % (51)	> − 90 % (>100)	−81 % (55)

cases, high fidelity surrogates outperform low fidelity surrogates in terms of the median RMSE.

The discrepancy function based MFS frameworks with a regression scalar gave more accurate surrogates in terms of median RMSE, especially for small sample size ratio as in the previous example. Figure 11 show that BDR and SDR frameworks significantly outperformed the other frameworks for all sample size ratios. This shows that the effect of applying full Bayesian treatment was minimal. The frameworks using a regression scalar, BDR and SDR, significantly outperformed the rest, and the Bayesian treatment did not matter. Similarly, Fig. 11e and f show that there was no noticeable difference between Bayesian and simple frameworks using calibration. For this example, the effect of a regression scalar ρ was significant since the other discrepancy based frameworks with $\rho = 1$, BD and SD, were significantly outperformed by BDR and SDR. For the frameworks using calibration, there was little effect of having a regression scalar ρ. This is because the same scaling effect can be achieved through calibration of H_u and H_l. Note that the median RMSE behaviors of cost ratio of 30 from Fig. 11 and for cost ratio of 10 were similar. The median RMSE of cost ratio of 10 is given in Fig. 14 in Appendix B.

Table 6 summarizes the results for all the cases in Table 4. The performance of BDR and SDR were significantly better than other frameworks using calibrations. The median RMSE improvements were higher than cost savings but still the effect of cost saving was significant. The effect of having a regression scalar was significant while applying full Bayesian approach had little effect for the discrepancy function based frameworks. The borehole function is a simple monotonically increasing function in the domain of input variables. For a simple function, discrepancy based approaches were better than frameworks using calibration and gains of applying Bayesian framework are limited.

4.3 Predicting the performance of the frameworks based on PRESS

The previous sections showed the performance of the frameworks based on the median of randomly generated 100 DOEs. We have observed, however, that even if one MFS is superior in most DOEs, it can perform poorly for others. Thus a performance prediction for a given problem and DOE would help to choose a proper MFS framework. Since *PRESS* is considered as a good performance predictor for surrogate models (e.g. Viana et al. 2009; Viana and Haftka 2009), we examined if *PRESS* can predict the performance of MFS frameworks. We selected sample ratios and predicted the ranking of all the frameworks based on their *PRESS*. Then we evaluated the predictability by comparing the predicted best framework to actual best framework based on RMSEs for each DOE. We counted how many cases out of the 100 DOEs *PRESS* made right predictions and obtained the success rate in those two measurements.

Table 7 shows the performance of *PRESS* for the Hartmann 6 function example. 10/560 and 20/240 were

Table 7 The numbers of cases out of 100: the predicted best/worst framework is among actual best 1, 2 and 3 (Hartmann 6 function example, 9 possible frameworks)

		Correct best/worst	Within best/worst 2	Within best/worst 3
10/560	Best	15	26	43
	Worst	28	49	60
20/240	Best	16	36	49
	Worst	20	47	62

Table 8 The numbers of cases out of 100: the predicted best/worst framework is among actual best 1, 2 and 3 (Borehole function example)

		Correct best/worst 1	Correct best/worst 2	Correct best/worst 3
5/50	Best	100	–	–
	Worst	45	72	78

selected since they were the optimal sample size ratio for the total budget of 28H and the cost ratio of 30 for BCDR and BDR, respectively. 10-fold cross validation errors were used to predict the ranks of the frameworks. We predicted the best and worst frameworks for a given DOE based on *PRESS*. We checked if the selected best or worst frameworks are within the actual best 1, 2 and 3 or worst 1, 2 and 3 frameworks in terms of RMSE (out of the 9 frameworks investigated here). For 10/560, *PRESS* predicted the actual best framework for only 15 DOEs out of 100 DOEs. Note that the cases for best 2 include the cases for best 1. Table 7 shows that the rate of making right prediction based on *PRESS* is not high. The chance that the predicted best framework is within the actual best 3 is 43 % and the chance that the predicted worst framework is within the actual worst 3 is 60 %. Also there is little difference between the two cases of 10/560 and 20/240. The performance of the frameworks for the sample size ratio 10/560 and 20/240 were competitive except SD, SDR and BD based on Fig. 9. The result tells us that *PRESS* could not rank the frameworks accurately.

Table 8 shows the predictive performance of *PRESS* for the borehole function example. We considered 5/50 for the total budget of 10H and the cost ratio of 10. As Fig. 11 shows, BDR and SDR significantly outperform the other frameworks and the *PRESS* performs much better than the Hartmann 6 function example.

5 Concluding remarks

In this paper, we attempted to provide insight on three Bayesian MFS frameworks and the corresponding simple frameworks based on approaches using 1) an additive discrepancy function, 2) calibration of low fidelity simulations, and 3) a comprehensive approach using both. We also examined the effect of the regression scalar ρ applied for MFS frameworks. The frameworks were examined with a 6D Hartmann 6 algebraic function and a 3D Borehole physical function.

The MFS frameworks were more potent for reducing computational cost rather than improving accuracy, especially for the Hartmann 6 function. The phenomenon was more obvious for the Hartmann 6 function than the borehole function. For the Hartmann 6 function, the discrepancy based MFS framework could build MFS surrogates equivalent to Kriging surrogates based on only high fidelity samples in terms of median RMSE with 14 % of the computational cost for the single fidelity Kriging surrogate (equivalent to 86 % cost saving). The maximum accuracy improvement over high fidelity Kriging surrogates is 51 % reduction in median RMSE. For the borehole function, the accuracy improvement is better than computational cost saving but the difference is marginal compared to the huge difference between the corresponding quantities for the Hartmann 6 function. The Bayesian discrepancy framework performed generally the best. However, performance of the framework deteriorated rapidly when the number of available high fidelity samples was low. In contrast, for the Hartmann 6 function, the Bayesian calibration framework performed reliably well for the case of few high fidelity samples and outperformed the Bayesian discrepancy framework. Since the discrepancy between the low and high fidelity functions was much more complex for the Hartmann 6 function than the borehole function, the discrepancy function prediction with few high fidelity samples had huge error for the Hartmann 6 function. Using a regression scalar ρ was particularly beneficial for the discrepancy based frameworks. For the Borehole function example, the presence of the scalar led to significant difference in the quality of an MFS and there is little benefit of using Bayesian framework.

We also found substantial differences in performance between frameworks for different DOEs, which raised the question whether *PRESS* based on cross validation errors can help us choose the best framework. We found that the usefulness of *PRESS* is limited. It is not as helpful as it is for single fidelity surrogates.

Finally, the Bayesian frameworks were compared to the simple frameworks using Kriging surrogates. One of the advantages of the simple frameworks, is that they can be much more readily implemented with existing surrogates. This may compensate for the relatively small advantage observed from the Bayesian frameworks.

This paper focuses on predicting the response of a high fidelity model with aid of a single low fidelity model using the MFS frameworks. However, high fidelity models can hardly be perfect, a higher fidelity model may need to be considered to measure the error of the high fidelity model. Alternatively, only a general knowledge of the magnitude of the error in the high-fidelity model is available. The knowledge can be useful to avoid trying to reduce the error of the MFS much below the high fidelity error.

Acknowledgments This work is supported by the U.S. Department of Energy, National Nuclear Security Administration, Advanced Simulation and Computing Program, as a Cooperative Agreement under the Predictive Science Academic Alliance Program, under Contract No. DE-NA0002378.

Appendix A: Statistical study of the 1-D function with 100 DOEs

The 1D function examples presented in the main text were selected to illustrate differences between frameworks because they exhibit distinctive differences. The Bayesian discrepancy function gave a significantly better prediction than the simple prediction and the Bayesian comprehensive framework gave very different calibration results than the other frameworks using calibration. The results were observed from the selected DOEs from randomly generated 100 DOEs. In this section, we show statistical representation for all 100 DOEs. The same trend functions and parameter bounds were used for fitting MFSs. For generating samples, we intendedly increased randomness by allowing a small number of iterations for LHS for initial low and high fidelity samples to see various cases.

Figure 12a presents the 100 RMSEs of the discrepancy based frameworks, SDR and BDR in the form of a boxplot. The center red line indicates the median (50 %) and the bottom and top of boxes are lower (25 %) and upper (75 %) quartiles of 100 RMSEs. The default distances of upper and lower whiskers between the upper and lower quartiles are $1.5w$ where w is the inter quartile distance which is the distance between upper and lower quartiles. If maximum or minimum samples are within the default bounds, whiskers are adjusted. Samples out of the default bounds are considered as outliers and they indicated with red crosses. The Bayesian discrepancy framework significantly outperforms the simple discrepancy framework statistically. The median RMSEs of BDR and SDR are 3.5 and 0.6, respectively. The correlation coefficient of the RMSEs of the two frameworks is 0.25 which is weak that one bad DOE for one framework may be a good DOE for the other. However, the mean and standard deviation of BDR are significantly smaller than SDR that SDR is better than BDR for a few DOEs with negligible difference. The means of regression scalar ρ of SDR and BDR are 0.54 and 1.92, respectively. That tells the ways of estimating ρ are responsible for the difference. There was weak correlation between RMSEs of the two frameworks but the worst DOEs for BDR are also bad DOEs for SDR RMSEs. The worst and the second worst RMSEs of BDR are 6.7 and 4.5 and the corresponding RSMEs of SDR are respectively 5.8 and 4.3.

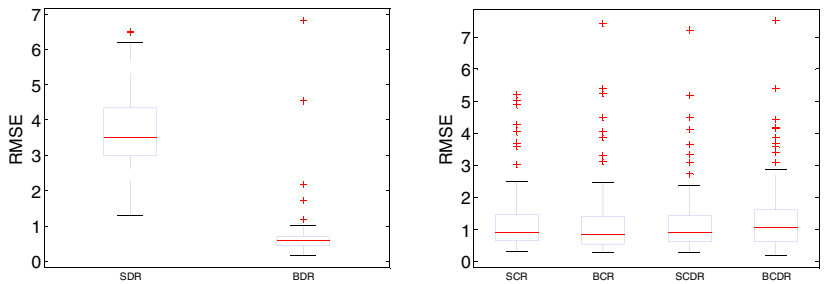
Table 9 Correlation coefficients between 100 RMSEs of the four frameworks

	BCR	SCDR	BCDR
SCR	0.9	0.9	0.8
BCR		0.9	0.9
SCDR			0.9

Figure 12b shows box plots of the frameworks using calibration, SCR, BCR, SCDR and BCDR. In terms of median RMSE, all frameworks show similar performance. Table 9 shows the correlation coefficients between RMSEs of the four frameworks. Unlike the previous discrepancy frameworks, they have very strong correlations. That means that a good DOE for one is highly likely to be a good DOE for the others that a good DOE is a necessary condition for constructing a good MFS. We searched for a DOE that was good for a framework and bad for another, but we could not find a single such DOE from the 100 DOEs.

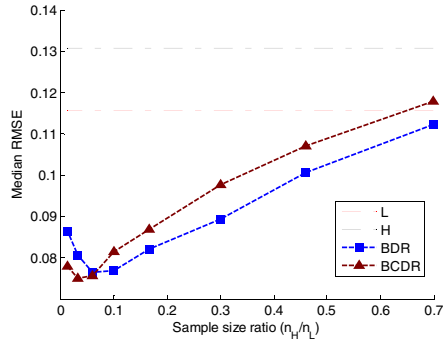
Appendix B: Median RMSEs for Different Sample Size Ratios

In the previous example section, only the median RMSEs for cost ratio of 30 were presented for both the Hartmann 6 function example and the borehole function example since there is no noticeable difference in the behavior between cost ratios of 30 and 10. In this appendix, the median RMSEs for cost ratio of 10 are presented in Fig. 13 for the Hartmann 6 function example and in Fig. 14 for the borehole function example.

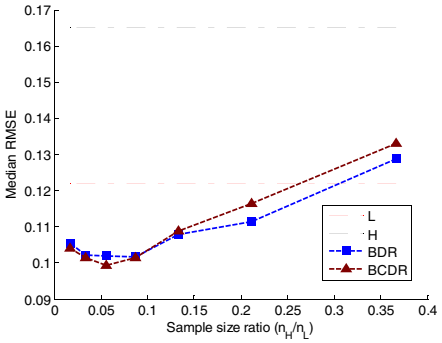


(a) Frameworks using a discrepancy function (b) Frameworks using calibration

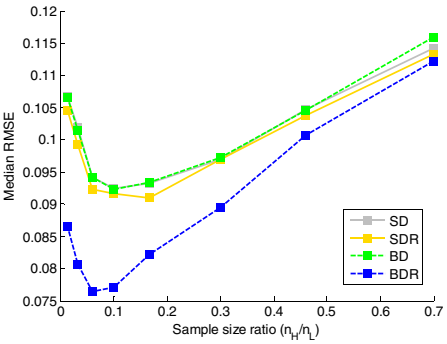
Fig. 12 Performance variations for 100 DOEs. **a** Frameworks using a discrepancy function. **b** Frameworks using calibration



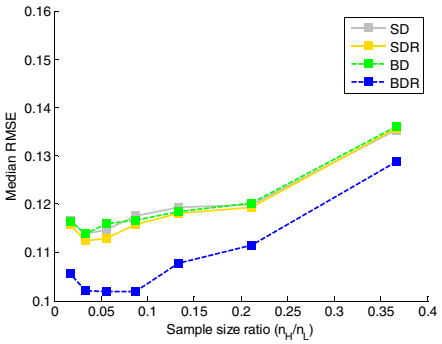
(a) Best frameworks for 56H total budget



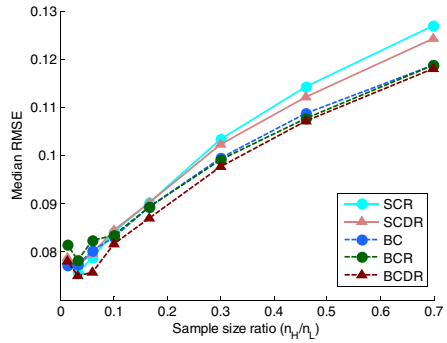
(b) Best frameworks for 28H total budget



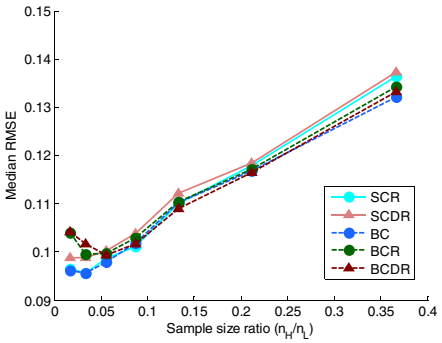
(c) Discrepancy based frameworks for 56H total budget



(d) Discrepancy based frameworks for 28H total budget

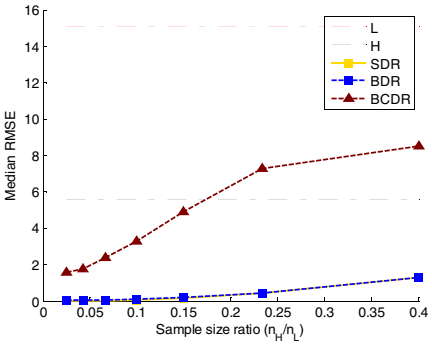


(e) Frameworks using calibration for 56H total budget

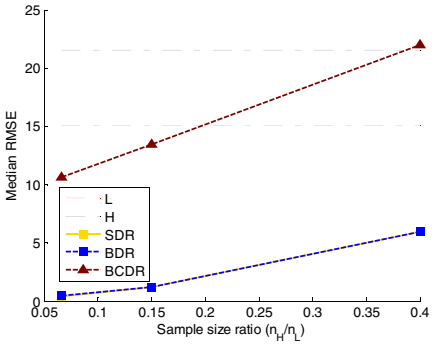


(f) Frameworks using calibration for 28H total budget

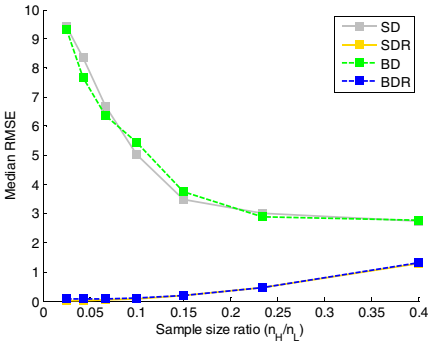
Fig. 13 Median RMSEs for different sample size ratios and cost ratio of 10. **a** Best frameworks for 56H total budget. **b** Best frameworks for 28H total budget. **c** Discrepancy based frameworks for 56H total budget. **d** Discrepancy based frameworks for 28H total budget. **e** Frameworks using calibration for 56H total budget. **f** Frameworks using calibration for 28H total budget



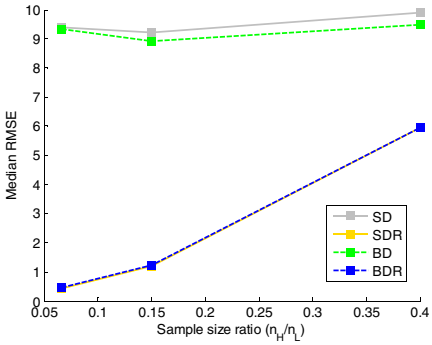
(a) Best frameworks for 10H total budget



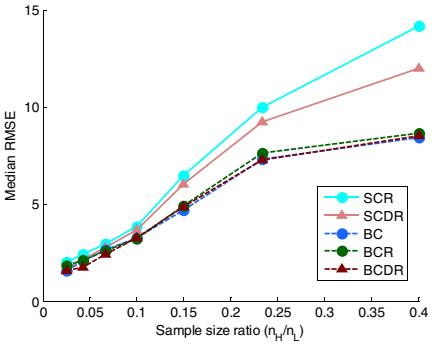
(b) Best frameworks for 5H total budget



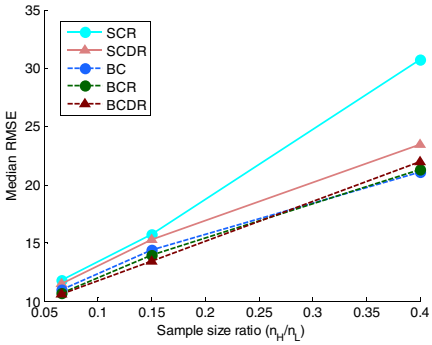
(c) Discrepancy based frameworks for 10H total budget



(d) Discrepancy based frameworks for 5H total budget



(e) Frameworks using calibration for 10H total budget



(f) Frameworks using calibration for 5H total budget

Fig. 14 Median RMSEs for different sample size ratios and cost ratio of 10. **a** Best frameworks for 10H total budget. **b** Best frameworks for 5H total budget. **c** Discrepancy based frameworks for 10H total budget. **d** Discrepancy based frameworks for 5H total budget. **e** Frameworks using calibration for 10H total budget. **f** Frameworks using calibration for 5H total budget

References

Acar E, Rais-Rohani M (2009) Ensemble of metamodels with optimized weight factors. Struct Multidiscip Optim 37(3):279–294

Balabanov V, Haftka RT, Grossman B, Mason WH, Watson LT (1998) Multifidelity response surface model for HSCT wing bending material weight. In Proceedings of 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization

Bayarri MJ, Berger JO, Paulo R, Sacks J, Cafeo JA, Cavendish J, Tu J (2007) A framework for validation of computer models. Technometrics 49:138–154

Coppe A, Pais MJ, Haftka RT, Kim NH (2012) Using a simple crack growth model in predicting remaining useful life. J Aircr 49(6): 1965–1973

Ellis MW, Mathews EH (2001) A new simplified thermal design tool for architects. Build Environ 36(9):1009–1021

Fischer CC, Grandhi RV (2014) Utilizing an adjustment factor to scale between multiple fidelities within a design process: a stepping stone to dialable fidelity design. In 16th AIAA Non-Deterministic Approaches Conference

Fischer CC, Grandhi RV (2015) A surrogate-based adjustment factor approach to multi-fidelity design optimization. In 17th AIAA Non-Deterministic Approaches Conference

Forrester AI, Sóbester A, Keane AJ (2007) Multi-fidelity optimization via surrogate modelling. Proc R Soc A 463(2088):3251–3269

Han Z, Zimmerman R, Görtz S (2012) Alternative Cokriging method for variable-fidelity surrogate modeling. AIAA J 50(5):1205–1210

Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD (2004) Combining field data and computer simulations for calibration and prediction. SIAM J Sci Comput 26(2):448–466

Jin R, Chen W, Sudjianto A (2005) An efficient algorithm for constructing optimal design of computer experiments. J Stat Plann Inference 134(1):268–287

Kennedy MC, O'Hagan A (2001). Supplementary details on Bayesian calibration of computer models. Internal Report. URL http://www.shef.ac.uk/~st1ao/ps/calsup.ps

Kennedy MC, O'Hagan A (2000) Predicting the output from a complex computer code when fast approximations are available. Biometrika 87(1):1–13

Kennedy MC, O'Hagan A (2001b) Bayesian calibration of computer models. J R Stat Soc Ser B (Stat Methodol) 63(3):425–464

Knill DL, Giunta AA, Baker CA, Grossman B, Mason WH, Haftka RT, Watson LT (1999) Response surface models combining linear and Euler aerodynamics for supersonic transport design. J Aircr 36(1): 75–86

Kosonen R, Shemeikka J (1997) The use of a simple simulation tool for energy analysis. VTT Building Technology

Kuya Y, Takeda K, Zhang X, Forrester AIJ (2011) Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. AIAA J 49(2):289–298

Le Gratiet L (2013) Multi-fidelity Gaussian process regression for computer experiments (Doctoral dissertation, Université Paris-Diderot-Paris VII)

Lee S, Youn BD, Sodano HA (2008) Computer model calibration and design comparison on piezoelectric energy harvester. In Proc. 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference (Victoria)

Lophaven SN, Nielsen HB, Søndergaard J (2002) DACE-A Matlab Kriging toolbox, version 2.0

Martin JD, Simpson TW (2005) Use of kriging models to approximate deterministic computer models. AIAA J 43(4):853–863

Mason BH, Haftka RT, Johnson ER, Farley GL (1998) Variable complexity design of composite fuselage frames by response surface techniques. Thin-Walled Struct 32(4):235–261

McFarland J, Mahadevan S, Romero V, Swiler L (2008) Calibration and uncertainty analysis for computer simulations with multivariate output. AIAA J 46(5):1253–1265

Morris MD, Mitchell TJ, Ylvisaker D (1993) Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. Technometrics 35(3):243–255

O'Hagan A (1992) Some Bayesian numerical analysis. Bayesian Stat 4(345–363):4–2

Owen AK, Daugherty A, Garrard D, Reynolds HC, Wright RD (1998) A parametric starting study of an axial-centrifugal gas turbine engine using a one-dimensional dynamic engine model and comparisons to experimental results: part 2—simulation calibration and trade-off study. In ASME 1998 International Gas Turbine and Aeroengine Congress and Exhibition (pp. V002T02A012-V002T02A012). Am Soc Mech Eng

Prudencio EE, Schulz KW (2012) The parallel C++ statistical library 'QUESO': Quantification of Uncertainty for Estimation, Simulation and Optimization. In Euro-Par 2011: Parallel Processing Workshops (pp. 398–407). Springer Berlin Heidelberg

Qian PZ, Wu CJ (2008) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics 50(2): 192–204

Rasmussen CE (2004) Gaussian processes in machine learning. In Advanced lectures on machine learning (pp. 63–71). Springer Berlin Heidelberg

Ryu JS, Kim MS, Cha KJ, Lee TH, Choi DH (2002) Kriging interpolation methods in geostatistics and DACE model. KSME Int J 16(5):619–632

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Stat Sci 409–423

Sanchez E, Pintos S, Queipo NV (2008) Toward an optimal ensemble of kernel-based approximations with engineering applications. Struct Multidiscip Optim 36(3):247–261

Viana FA, Haftka RT (2009) Cross validation can estimate how well prediction variance correlates with error. AIAA J 47(9):2266–2270

Viana FA, Haftka RT, Steffen V Jr (2009) Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Struct Multidiscip Optim 39(4):439–457

Xiong S, Qian PZ, Wu CJ (2013) Sequential design and analysis of high-accuracy and low-accuracy computer codes. Technometrics 55(1): 37–46

Yoo MY, Choi JH (2013) Probabilistic calibration of computer model and application to reliability analysis of elasto-plastic insertion problem. Trans Korean Soc Mech Eng A 37(9):1133–1140

Zheng L, Hedrick TL, Mittal R (2013) A multi-fidelity modelling approach for evaluation and optimization of wing stroke aerodynamics in flapping flight. J Fluid Mech 721:118–154